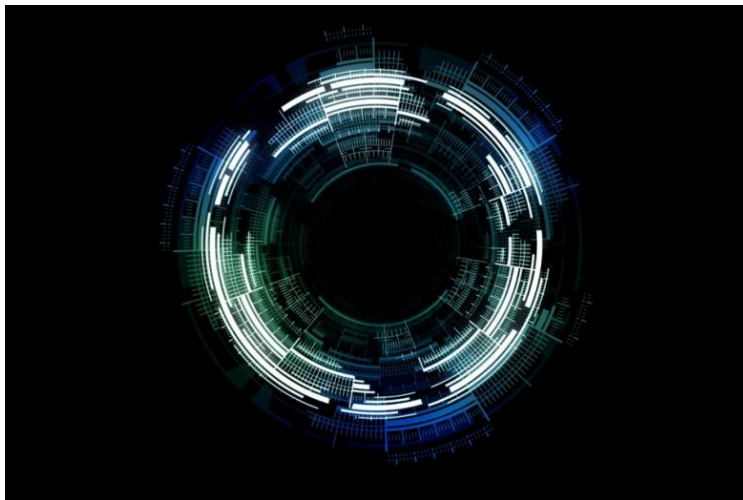




More bit than bio – communicating and discussing Whole Brain Emulation, Mind Uploading & Superintelligence using thought experiments.

Filipe Luig

Doutoramento em Ensino e Divulgação Científica
Especialidade – Divulgação da Ciência
2020

Orientador

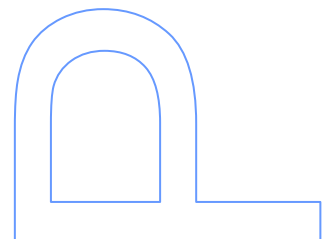
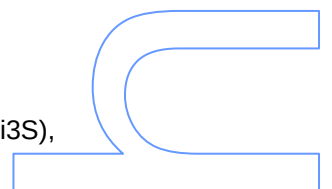
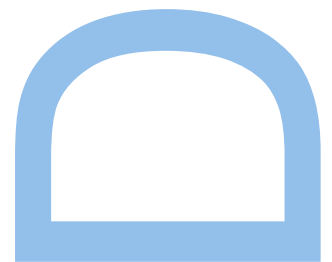
João Relvas, Investigador Principal, Instituto de Investigação e Inovação e Saúde (i3S),
Universidade do Porto, Professor Associado (convidado) FMUP

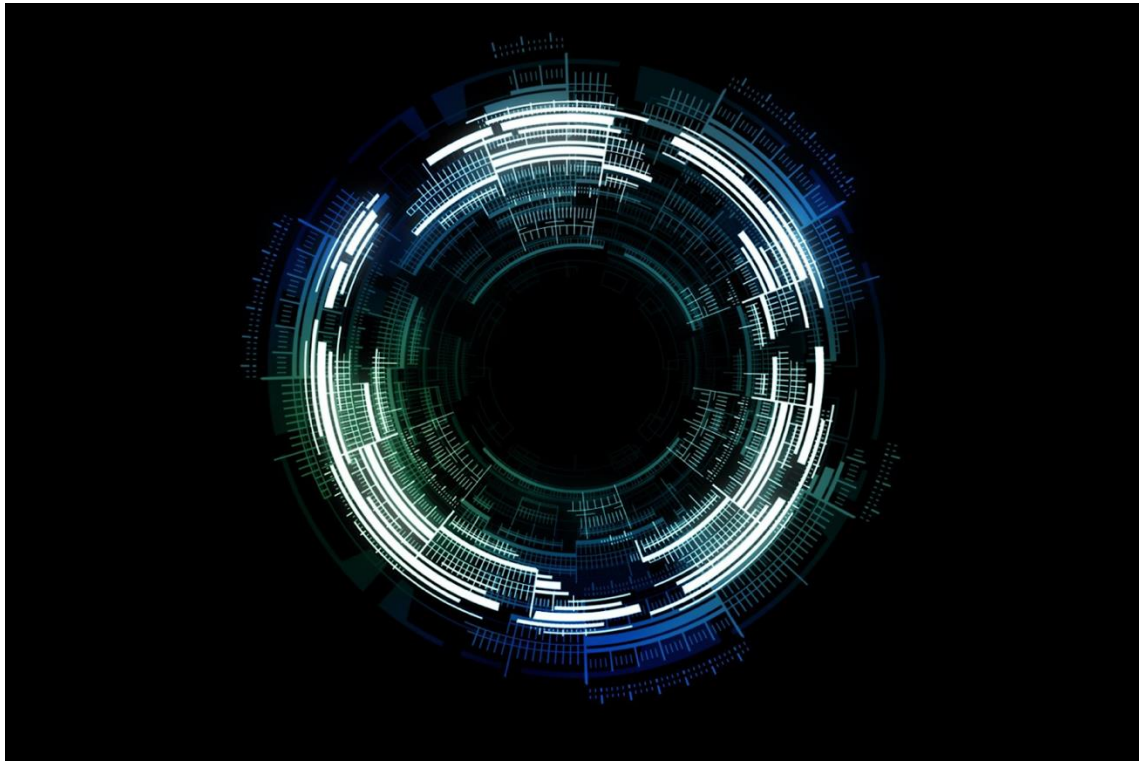
Coorientador

Gonçalo Marcelo, Investigador Contratado, CECH, Universidade de Coimbra

Coorientador

Luís Calafate, Professor Auxiliar
Departamento de Biologia
Faculdade de Ciências da Universidade do Porto (FCUP)





More bit than bio – communicating and discussing Whole Brain Emulation, Mind Uploading & Superintelligence using thought experiments.

Filipe Luig

PhD 2020

“Two possibilities exist: either we are alone in the Universe or we are not. Both are equally terrifying.”

— **Arthur C. Clarke**

Acknowledgments

Very special thanks to all the experts interviewed for this thesis, namely: Randal Koene, Anders Sandberg, David Chalmers, Kenneth Hayworth, Nicholas Agar, Tom Koch, James Barrat, Robin Hanson, Eric Cattell, David Pearce and Richard Loosemore (*ordered chronologically according to the dates of the interviews*).

Their contribution was essential to turn this work into a balanced and rich thesis given the variety of interesting views expressed. Each and every interview was a source of knowledge and a pleasure! Thanks a lot!

Thanks also to the supervisors of this thesis, namely to João Relvas, Gonçalo Marcelo and Luís Calafate for being always approachable and insightful as well as particularly patient with me!

Special Thanks to Gonçalo for *playing the devil's advocate* so well which also enabled some *hot and colorful* discussions.

Thanks to Filipa Ribeiro for helping me out with the analytical assessment of the interviews.

Thanks to Professor João Paiva for encouraging me to pursue these controversial fields and helping me out in *the early days* of this PhD.

Thanks to the *Faculdade de Ciências da Universidade do Porto*, and namely to Professor Clara Vasconcelos and the Scientific Committee of the doctoral programme in Science Teaching and Communication (PDEDC) for making it possible for me to write this thesis about a fascinating subject that affects all of us human beings and makes our future an even more thrilling discussion.

Abstract & Keywords

This thesis revolves on the topics of whole brain emulation, mind uploading and superintelligence that are ingrained with specifically conceived TEs to each of the chapters / topics, creating an environment to discuss and communicate scientific topics of the utmost importance for our common future.

Three thought experiments were designed for this thesis in the context of mind uploading and superintelligence. They are presented in a first chapter dedicated to clarify the potential and applicability of thought experiments communicating fundamental ideas in these scientific and technological fields, because the controversial (and often speculative) nature of these topics call for an out-of-the-box approach to be better analyzed and discussed. In the second chapter, whole brain emulation is thoroughly explored in its technical feasibility, paving the way for a metaphysical approach of mind uploading in the third chapter. Personal identity is, with the help of these thought experiments (*TE: iterating human predictor response*), scrutinized given that its current conceptual form does not provide a totally satisfactory explanation for several cases that come to light through mind uploading possible scenarios. Consciousness is then also analyzed in the light of existentialism and through the aforementioned thought experiments (*TE: the delusion of reality*), exploring the possibility of upgrading it, just as intelligence is predicted to be upgraded in the near future. It is argued that mind uploading could enable consciousness to perpetuate and that this purpose also counts as a survival backup of humankind making of mind uploading the means to maintain human thought *alive* throughout time. The fourth chapter explores the impact and possible outcomes superintelligence entails. Here, another of these thought experiments (*TE: iterating human intelligence*) is applied as an evolutionary algorithm approach to argue about the competitive balance humankind might have in order to avoid artificial superintelligence to either overcome or even subdue our species. The last chapter of this thesis is an analysis of a number of semi structured interviews conducted with experts in fields such as neuroscience, computing science and philosophy, among others. These were carried out to bring valuable input on the topics discussed throughout the thesis.

Keywords: Thought Experiments; Mind Uploading; Superintelligence; Whole Brain Emulation; Evolutionary Algorithm; Personal Identity; Consciousness.

Resumo

Esta tese gira em torno dos campos da emulação o cérebro, do *mind uploading* e da superinteligência que estão enraizados em TEs especificamente concebidos para cada um desses temas, contribuindo para a criação de um ambiente de discussão e comunicação de temas científicos da maior importância para o nosso futuro em comum.

Foram projetados três *thought experiments* no contexto do *mind uploading* e da superinteligência. Numa primeira instância são apresentados no primeiro capítulo dedicado a esclarecer o potencial e a aplicabilidade de *thought experiments* na comunicação de ideias fundamentais nesses campos científicos e tecnológicos, porque a natureza controversa (e muitas vezes especulativa) desses temas exige uma abordagem alternativa para uma melhor análise e discussão. No segundo capítulo é explorada a emulação do cérebro e a sua viabilidade técnica abrindo caminho para uma abordagem metafísica do *mind uploading* no terceiro capítulo. A identidade pessoal é, com a ajuda desses *thought experiments* (TE: *iterating human predictor response*), analisada, uma vez que o seu conceito atual não fornece uma explicação totalmente satisfatória para vários casos expostos nos possíveis cenários associados ao *mind uploading*. A consciência é então também analisada através de *thought experiments* (TE: *the delusion of reality*), e também à luz do existencialismo, explorando a possibilidade de um seu upgrade poder ocorrer acompanhando o incremento expectável de inteligência como previsto e visionado num futuro próximo. É também argumentado que o *mind uploading* possa permitir a perpetuação da consciência e também que essa perpetuação contribua como *backup* de sobrevivência da humanidade fazendo do *mind uploading* uma forma de manter o pensamento humano vivo ao longo do tempo. O quarto capítulo explora o impacto e possíveis consequências que uma superinteligência acarreta. Aqui, outro desses *thought experiments* (TE: *iterating human intelligence*) é aplicado como uma abordagem de algoritmo evolutivo argumentando em prole dum equilíbrio competitivo que a humanidade possa ter, a fim de evitar a sua possível subjogação por uma superinteligência artificial. O último capítulo assenta na análise de entrevistas semiestruturadas realizadas com especialistas em áreas como neurociência, ciências da computação e filosofia que foram realizadas com o intuito de contribuir de forma crítica na análise dos temas discutidos na tese.

Table of Contents

<u>Acknowledgments</u>	IV
<u>Abstract & Keywords</u>	V
<u>Resumo</u>	VI
<u>List of Figures</u>	XVII
<u>List of Tables and Charts</u>	XX
<u>Abbreviations</u>	XXII
<u>INTRODUCTION</u>	1
<u>General Objectives of the thesis</u>	6
<u>Methodology used in the different Parts of the thesis</u>	8
<u>Part I _ Thought Experiments</u>	9
<u>Ethical Disclaimer</u>	10
<u>Part II _ Whole Brain Emulation</u>	11
<u>Part III _ Mind Uploading</u>	12
<u>Part IV _ Superintelligence</u>	12
<u>Part V _ Semi-Structured Interviews – Analysis</u>	13
<u>Final Notes</u>	13
<u>Part I _ THOUGHT EXPERIMENTS</u>	15
<u>1 _ Thought Experiments (<i>Gedankenexperimente</i>)</u>	16
<u>1.1 _ intro to thought experiments</u>	16
<u>1.2 _ relation between TEs and empirical experiments</u>	20
<u>1.3 _ some classical thought experiments</u>	23

<u>1.4 some rules regarding the conception of a thought experiment</u>	27
<u>1.5 conceiving the thought experiment</u>	28
<u>2 Innovative technologies as a scientific support to mind-uploading</u>	31
<u>2.1 artificial general intelligence</u>	32
<u>2.2 intelligence explosion</u>	32
<u>2.3 molecular nanotechnology</u>	33
<u>2.4 whole brain emulation and mind uploading</u>	35
<u>2.5 the consciousness conundrum</u>	37
<u>2.6 concluding remarks</u>	38
<u>3 Mind Uploading / Superintelligence Thought Experiments</u>	38
<u>3.a TE: iterating human predictor response</u>	38
<u>3.b TE: iterating homo sapiens intelligence</u>	41
<u>3.c TE: the delusion of reality</u>	42
<u>Part II WHOLE BRAIN EMULATION</u>	45
<u>1 Intro</u>	46
<u>2 WBE Objectives</u>	50
<u>3 Technical Requirements</u>	51
<u>3.1 Scanning the brain (and the amount of neurobiological information necessary)</u>	51
<u>3.2 Translating the scan into a model (and the extent to which the reimplementation would have to be similar)</u>	54
<u>3.2.1 The Connectome</u>	59

<u>3.3 Running that model on a computer (and the criteria used in order to accurately evaluate the success rate of that emulation)</u>	60
<u>3.3.1 Non-organicism</u>	63
<u>4 Analytical Inferences</u>	66
<u>4.1 Connectomics @Plot</u>	67
<u>4.2 Brain Scanning (Techniques & Technologies) @Plot</u>	68
<u>4.3 Computing Power @Plot</u>	69
<u>4.4 Non-Organicism @Plot</u>	74
<u>4.5 AGI (Artificial General Intelligence) @Plot</u>	74
<u>4.6 @Plots</u>	75
<u>5 Preliminary Concluding Remarks A</u>	77
<u>6 Analogical Linkage (using paradigms towards WBE)</u>	78
<u>6. §1 Biomimicry towards AGI (and WBE)</u>	78
<u>6. §2 Exponential Progress Rate</u>	84
<u>6. §3 Classical - Quantum Transition</u>	86
<u>6. §4 Synergy (and connectomics)</u>	89
<u>6. §5 Brain Preservation (and postponing the reset)</u>	92
<u>7 Preliminary Concluding Remarks B</u>	94
<u>Part III MIND UPLOADING</u>	97
<u>1 Intro / Technical background</u>	98
<u>2 Approaches to Mind Uploading</u>	101

<u>3</u>	<u>Consciousness Conundrum</u>	104
3.1	<u>Uploading Consciousness</u>	106
3.2	<u>Debating the Hard Problem</u>	111
<u>4</u>	<u>Personal Identity</u>	113
4.1	<u>Main Personal Identity Theories</u>	117
4.2	<u>The neo-ship of Theseus</u>	118
4.3	<u>Branching Identity</u>	119
4.4	<u>Iterating Human Predictor Response (CaseScenario 1 to 3)-I</u>	122
4.5	<u>Iterating Human Predictor Response (CaseScenario 4) –II</u>	126
4.6	<u>Social Construction of Identity</u>	128
4.7	<u>Metaphysical Equivalence of Personal Identity in Uploading Scenarios or Scan and Copy Vs Gradual in-place Replacement</u>	129
4.8	<u>The Argument of Space (distance and spatial displacement)</u>	131
4.9	<u>The Argument of Time (the rate and temporal displacement)</u>	132
<u>5</u>	<u>Metaphysical Purposes of MU</u>	134
5.1	<u>Intro</u>	134
5.2	<u>To Perpetuate Consciousness</u>	135
5.3	<u>The Delusion of Reality (in the light of existentialism) - I</u>	136
5.4	<u>The Delusion of Reality (in the light of existentialism) - II</u>	139
5.5	<u>Essence and Existence</u>	142
<u>6</u>	<u>Critics and Criticism</u>	144
6.1	<u>Frequently Made Objections</u>	145

- [6.1.1 Only quantum computing would be able to emulate neural processes \(hereby dismissing “regular computing power”\)](#) 145
- [6.1.2 MU is such a large undertaking it deems itself unachievable](#)
- [6.1.3 “We lack a kind of “neural Rosetta code” which blocks us from ever understanding how the brain processes information. It is like trying to decode a CD without having the CD player’s software” \(Bonilla, 2017\)](#) 145
- [6.1.4 “Minds only operate properly in the context of an associated body and its sensory-motor interactions with the physical world.” \(Lakoff, 1980\)](#) 146
- [6.1.5 “There is no reason to suppose an exact functional copy of the brain will produce similar phenomenological effects and also that the uploaded mind cannot entail the same self as the original brain.” \(Hauskeller, 2012\)](#) 148
- [6.1.6 “But even if the mind resulting from a particular brain emulation will indeed be qualitatively identical \(...\) which is an additional requirement.” \(Hauskeller, 2012\)](#) 150
- [6.1.7 “The ancient paradox known as the Ship of Theseus \(...\) The question then arises whether the ship is still the same ship or rather a different one.” \(Hauskeller, 2012\)](#) 151
- [6.1.8 Information idealism and the walking algorithms argument](#)
- [6.1.9 “Although, let’s face it, we don’t have the slightest clue how conscious experience comes about \(...\) it is rather unlikely that mind and self are directly produced by the brain, as is commonly assumed.” \(Hauskeller, 2012\)](#) 153
- [6.1.10 Regarding instantaneous destructive uploading, a recurrent objection has to do with the supposed distance to travel \(...\) from the original brain to a computer. \(Corabi, 2014\)](#) 153

6.1.11	<u>“The main reasons for being pessimistic have to do with reservations about positions that attribute strange and discontinuous motions (...) in order to claim that uploading is compatible with survival.” (Corabi & Schneider 2012)</u>	154
6.1.12	<u>Regarding gradual destructive uploading, a recurrent objection has to do with temporal discontinuity</u>	155
6.1.13	<u>“I combine the claim that (...) mind uploading will fail with the claim that comparatively little is gained (...) mind-uploading is prudentially irrational” (Agar, 2011)</u>	155
6.1.14	<u>“Consciousness is – to the best of our knowledge – a biological phenomenon that is unlikely to be entirely (or even largely) substrate-independent.” (Pigliucci, 2014)</u>	157
6.2	<u>Ethics</u>	160
6.3	<u>Approval and Condemnation of MU</u>	165
6.4	<u>MU Validation (The Terasem MU Experiment, Rothblatt, 2012)</u>	168
7	<u>Concluding Remarks</u>	170
7.1	<u>Reconstructive uploading</u>	171
7.2	<u>Reinventing Identity</u>	172
7.3	<u>Further Facts</u>	173
7.4	<u>Final Notes</u>	174
8	<u>Reasonable Inferences</u>	174
8. §1	<u>Split-Syndrome / Comatosis / Anaesthesia / Amnesia / Multi-Personality</u>	174
8. §2	<u>Qualia (and quantics)</u>	177
8. §3	<u>Measuring Consciousness</u>	181
8. §3.1	<u>Machine Consciousness</u>	184

<u>PART IV SUPERINTELLIGENCE</u>	185
<u>1 Intro</u>	186
<u>2 Singularity = Acceleration + Discontinuity + Superintelligence</u>	188
<u>2.1 Two conflicting visions and/or views (Bostrom vs Kurzweil)</u>	191
<u>3 Two Singularity Outcomes: IA (intelligence amplification) vs AI (artificial intelligence)</u>	194
<u>4 Intelligence</u>	196
<u>4.1 TE 3b iterating human intelligence</u>	199
<u>5 Artificial Intelligence (ANI-AGI-ASI)</u>	202
<u>5.1 Artificial Narrow Intelligence (ANI) / Artificial General Intelligence (AGI) / Artificial Super Intelligence (ASI)</u>	203
<u>5.2 The current state of AI</u>	204
<u>5.3 The AI Predictions Issue and (un)reliable experts</u>	211
<u>6 Intelligence Explosion</u>	214
<u>7 AI Safety (and risks of the Singularity)</u>	216
<u>7.1 Strategies to control (artificial) intelligence</u>	218
<u>8 Politics & Skepticism</u>	222
<u>9 Concluding Remarks</u>	224
<u>9.1 Conflicting views and different approaches I</u>	225
<u>9.2 The fear of a singleton (and of non-human consciousness?)</u>	226
<u>9.3 Conflicting views and different approaches II</u>	228
<u>9.4 Final notes</u>	229
<u>PART V SSI ANALYSIS</u>	231

<u>1 Intro and methodology</u>	232
<u>2 Objectives</u>	236
<u>3 Interview's Guide</u>	237
<u>4 SSI's Rationale</u>	240
<u>5 Content Analysis</u>	245
<u>5.1 Question I</u>	245
Regarding Whole Brain Emulation and Mind-Uploading	245
<u>5.1.1 Q1: Would another me be me? (or just a copy of myself)</u>	
<u>5.1.2 Analysis #1</u>	250
<u>5.1.3 Q2: What would represent the biggest risk in achieving WBE/MU?</u>	252
<u>5.1.4 Analysis #2</u>	257
<u>5.2 Question II</u>	260
Regarding Superintelligence and the Technological Singularity	
<u>5.2.1 Q1: If (When) it happens (and we're still humans) will we even know it (be aware of)? Could this mean we'd be some kind of a cyber sapiens by this time?</u>	260
<u>5.2.2 Analysis #3</u>	265
<u>5.2.3 Q2: How possible would it be for a Singleton to rule the world</u>	269
<u>5.2.4 Analysis #4</u>	272
<u>5.3 Question III</u>	274
Regarding Transhumanism and Posthumanism	
<u>5.3.1 Q1: Being human will be different in a Post-human Era. Could this mean we would be more bit than bio by then?</u>	274

<u>5.3.2 Analysis #5</u>	279
<u>5.3.3 Q2: Do you think politics in a few years will be divided between bioconservatives vs transhumanists?</u>	282
<u>5.3.4 Analysis #6</u>	287
<u>5.4 Question IV</u>	291
Regarding Superintelligence and mind uploading	
<u>5.4.1 Q1: Would it make sense to iterate human intelligence via MU to augment it to cope with AI's progress?</u>	291
<u>5.4.2 Analysis #7</u>	297
<u>5.4.2.1 About the TE3b - iterating human intelligence</u>	301
<u>5.4.3 Q2: Could MU prove that two different persons sharing an identical blueprint would in fact be identical? If so, would this not shake our beliefs in issues such as personal identity?</u>	305
<u>5.4.4 Analysis #8</u>	307
<u>5.5 Question V</u>	310
Regarding Transcending Biology and Human Enhancements	
<u>5.5.1 Q1: Is it inevitably going to generate a dystopian society?</u>	
<u>5.5.2 Analysis #9</u>	312
<u>GENERAL CONCLUSIONS</u>	315
<u>TEs:</u>	315
<u>To reinvent identity</u>	317
<u>To keep the competitive balance</u>	317
<u>To perpetuate consciousness</u>	319
<u>WBE: The Feasibility Discussion</u>	320

<u>MU: Reinventing identity, hence reinventing consciousness (or the other way around)</u>	321
<u>SI: The difficulty and fear of dealing with the unknown</u>	321
<u>SSI-Analysis: Some uniformity among the interviewed experts</u>	322
<u>Limitations and recommendations for future research</u>	323
<u>Final note (Ethical Disclaimer)</u>	325
<u>REFERENCES</u> (+Webinars and Workshops)	327
<u>APPENDIX</u>	354
<u>APPENDIX A _ Request for Interview</u>	355
<u>APPENDIX B _ Candidates contacted for interview</u>	356
<u>APPENDIX C _ Interviewees' Presentation</u>	359
<u>APPENDIX D _ Interview with Randal KOENE</u>	365
<u>APPENDIX E _ Interview with David CHALMERS</u>	381
<u>APPENDIX F _ Interview with Anders SANDBERG</u>	384
<u>APPENDIX G _ Interview with Kenneth HAYWORTH</u>	402
<u>APPENDIX H _ Interview with Nick AGAR</u>	416
<u>APPENDIX I _ Interview with Tom KOCH</u>	430
<u>APPENDIX J _ Interview with James BARRAT</u>	450
<u>APPENDIX K _ Interview with Rick CATTELL</u>	467
<u>APPENDIX L _ Interview with Robin HANSON</u>	482
<u>APPENDIX M _ Interview with David PEARCE</u>	493
<u>APPENDIX N _ Interview with Richard LOOSEMORE</u>	504
<u>APPENDIX O _ Calculus for Plot A & Plot B (WBE feasibility Plot)</u>	522

List of Figures

- [Figure 01](#) – Concerning artificial intelligence (...): When will AGI override humankind and will it be beneficial to us? (Source: Max Tegmark, 2017) 14
- [Figure 02](#) – Spatial resolution of various brain scanning technologies versus temporal resolution, showing the differences between functional and structural applicabilities of them. (Source: Pfister, et al, Visualization in Connectomics, Mathematics and Visualization June 2012) 55
- [Figure 03](#) – Spatial resolution of Magnetic Resonance Imaging (MRI) for use in brain scanning, in microns. Since the y-axis is logarithmic and the curve is more or less linear, the resolution has improved more or less exponentially. (Kurzweil, 2012) However, even if this trend continues for the decades to come MRI still will not have the spatial resolution needed to see pieces of neuroanatomy relevant to WBE (including synapses). (Source: Kurzweil, 2012) 56
- [Figure 04](#) – Calculations per second (measured in floating-point operations per second) per constant \$1000 of hardware. Note that the y-axis is logarithmic, so the increasing slope indicates more than exponential growth. This spans back over a century and will continue until hardware requirements for WBE can be available on computers costing no more than what PCs cost today. (Source: Kurzweil, 2010) 69
- [Figure 05](#) – (Plot A: WBE feasibility regarding critical issues. (Luig, 2020) 76
- [Figure 06](#) – (Plot B): WBE feasibility regarding critical issues. (Luig, 2020) 77
- [Figure 07](#) – Screenshot taken from cover of nano-magazine issue March 2020. 80
- [Figure 08](#) – Screenshot taken from cover of nano-magazine issue December 2019. 81
- [Figure 09](#) – A device like the one in the study (right), and an electron microscope image showing the device's neuron-like arrangement of nanowires. (Source: Photo composite by Marc Roseboro/CNSI at UCLA) 82
- [Figure 10](#) – Screenshot taken from cover of nano-magazine issue December 2018 83

[Figure 11](#) – Spatial resolution of Magnetic Resonance Imaging (MRI) for use in brain scanning, in microns. Since the y-axis is logarithmic and the curve is more or less linear, the resolution has improved more or less exponentially. (Source: Kurzweil, 2005). 84

[Figure 12](#) – Exponential growth of supercomputing showing estimated timeline of reaching and surpassing human brain capacity in FLOPS. (Source: available on the internet) 86

[Figure 13](#) – On the left physical miniaturization of silicon devices is hitting the wall of its limit, which is coincident with the transition of the classical to the quantum age (on the right). (Source: Wilson, M. et al, *Nanotechnology – basic science and emerging technologies*, 2002) 87

[Figure 14](#) – On the left representation of classical Bit. On the right the quantum correspondent - Qubit. (Source: available on the internet) 88

[Figure 15](#) – Brain Projects' Timeline. (Source: available on the internet) 89

[Figure 16](#) – Screenshot taken from cover of nano-magazine issues Sep+Oct 2019 90

[Figure 17](#) – Screenshot taken from cover of nano-magazine issues Mar17 + Jan19 91

[Figure 18](#) – Screenshot taken Speculative Timeline during Hayworth's Talk. (Source: Hayworth, 2018) 94

[Figure 19](#) – Gradual replacement w/ neuroprosthetic devices for each brain piece. (Source: carboncopies 2018 / <http://www.minduploading.org/>) 102

[Figure 20](#) – Reconstruct a working artificial brain from the information in tiny brain slices. (source: carboncopies 2018 / <http://www.minduploading.org/>) 103

[Figure 21](#) – Predictive level of facts. Unique facts form a rather narrow strip between common and random facts; inside this strip, the curve declines almost linearly and doesn't change much. (Source: Alexey Turchin, *Theory and Protocol for Indirect Mind Uploading*, 2018) 125

[Figure 22](#) – (adapted from Linssen et al): Diagram relating embeddedness, embodiment, and substrate. (A) Present situation: the brain is contained in the body which is contained in the environment. Action potentials are routed between brain and body via nerve fibers (dashed line). (B) Situation for WBE: the substrate is no longer fixed to the body. Action potentials are electronically routed between the emulation and body (dashed line). (Source: Linssen et al, 2016) 147

[Figure 23](#) – Global NLP market size growth projection. (Source: KBV, Knowledge Based Value Market Research Report, 2019) 206

[Figure 24](#) – This linear plot shows microprocessor power measured in FLOPS over the last 25 years. (Source: Top500 Supercomputer Database) 207

[Figure 25](#) – This linear plot shows microprocessor power measured in clock speed, as the number of pulses per second (in Hertz) generated by an oscillator that sets the tempo for the processor. (Source: Top500 Supercomputer Database) 207

[Figure 26](#) – Exponential growth of dataset in the few years. (Source: <https://research.aimultiple.com/ai/>) 209

[Figure 27](#) – Venture capital/ startup funding growth in the last few years. (Source: statista charts, 2020) 210

[Figure 28](#) – Median estimate for human-level AI, graphed against date of prediction. (Source: Wang and Potter, MIRI, 2012) 213

[Figure 29](#) – Screenshot taken from cover of nano-magazine issue June 2017 230

[Figure 30 & Figure 31](#) – SSI-reprise I & II – show the pitch-kind of answers given by the interviewees to each specific issue in terms of probability. (Luig, 2020) 314

[Figure 32](#) – Connectomics, Brain-Scanning and Computing Power quite reliable, whereas AGI and Non-Organicism are quite uncertain regarding the path towards a WBE feasibility. (Luig, 2020) 523

List of Tables and Charts

Chart 01 : WBE and its main fields of interaction analyzed in the plots A and plot B (Figure 05 & Figure 06) below as they devise essentially in two big areas. The computing field with its own actors and the neuroscientific field, also with its own main topics of influence on WBE. (Source: Luig, 2020)	66
Table 01 – Accelerated progress in generations (Roco and Bainbridge 2006)	04
Table 02 – Terminology adopted by Max Tegmark in his book <i>Life 3.0, Being Human in the Artificial Intelligence Age</i> used for definitions that are sometimes hard to describe and many times misused and misapplied. (Source: Tegmark, M., 2017)	05
Table 03 – Storage demands (emulation only, human brain). (Source: brain emulation roadmap, Sandberg and Bostrom, 2008)	72
Table 04 – Processing demands (emulation only, human brain). (Source: brain emulation roadmap, Sandberg and Bostrom, 2008)	73
Table 05 – Link between the general objectives of the thesis and the questions and respective analysis posed by the interviews	237
Table 06 – Interview Guide	238
Table 07 – Overview and summary of questions and interviewees with their fields of expertise. (Luig, 2020)	243
Table 08 – Overview and summary of questions and interviewees with their fields of expertise. (Luig, 2020)	244
Table 09 – Selection of transcript excerpts regarding Question I – Q1	245
Table 10 – Selection of transcript excerpts regarding Question I – Q2	253
Table 11 – Selection of transcript excerpts regarding Question II – Q1	260
Table 12 – Selection of transcript excerpts regarding Question II – Q2	269

Table 13 – Selection of transcript excerpts regarding Question III – Q1	274
Table 14 – Selection of transcript excerpts regarding Question III – Q2	282
Table 15 – Selection of transcript excerpts regarding Question IV – Q1	291
Table 16 – Selection of transcript excerpts regarding Question IV – Q2	305
Table 17 – Selection of transcript excerpts regarding Question V – Q1	310
Table 18 – List of Experts contacted for interview. (Luig, 2020)	356
Table 19 – Calculus and critical values attributed for Plots A & B (WBE Feasibility)	522
Table 20 – Plot B <i>WBE Feasibility</i> data	524

Abbreviations

AI – Artificial Intelligence

AFM – Atomic Force Microscopy

AGI – Artificial General Intelligence

ANI – Artificial Narrow Intelligence

ASI – Artificial Super Intelligence

BCI – Brain Computer Interface

BPF - Brain Preservation Foundation

CToM - Computational Theory of Mind

CPS - Calculations Per Second

DID - Dissociative Identity Disorder

EEG – Electroencephalography

Em – Emulations (*as in uploads*)

FLOPS - Floating Operations Per Second

FMO - Frequently Made Objections

fMRI – functional Magnetic Resonance Imaging

HBP - Human Brain Project

HLAI – Human Level Artificial Intelligence

IA – Intelligence Augmentation (in reference to the IA Singularity)

IBM – Intelligence Business Machines

IITC – Information Integration Theory of Consciousness

IT – Information Technology

IQ – Intelligence Quocient

LoAR – Law of Accelerating Returns

MEG – Magneto Encephalography

MIRI – Machine Intelligence Research Institute (*former Singularity University*)

MIT – Massachusetts Institute of Technology

ML – Machine Learning

MU – Mind Uploading

MURG – Mind Uploading Research Group

NBIC – Nano-Bio-Info-Cognitive (*technologies*)

NLP – Natural Language Processing

PCI – Perturbational Complexity Index

PET - Positron Emission Tomography

RAM – Random Access Memory

RL – Reinforcement Learning

REM – Rapid Eye Movement (sleep)

Sci-Fi – Science Fiction

SET – Single Electron Transistor

SI – Superintelligence

SIM – Substrate Independent Mind

STM – Scanning Tunneling Microscope

TE – Thought Experiment

TIPS – Trillion Instructions Per Second

WBE – Whole Brain Emulation

Introduction

“I believe there is no deep difference between what can be achieved by a biological brain and what can be achieved by a computer. It, therefore, follows that computers can, in theory, emulate human intelligence — and exceed it.” Quote from a speech given by Hawking at the opening of the Leverhulme Centre of the Future of Intelligence, Cambridge, UK, October 2016.

More bit than bio. It is this perhaps intriguing *leitmotiv* that I choose to frame the work undertaken for this PhD thesis in these last few years. It begs a few words of explanation, and with them we shall start introducing the readers to this work.

This research took place in the context of the doctoral programme in Science Teaching and Communication (PDEDC) taking place at the Faculty of Sciences of the University of Porto. This is a vibrant and strongly interdisciplinary programme that always struck me as fostering the discussion of cutting-edge scientific topics, problems and solutions. In this context, this thesis, which specializes in the communication of science, puts forward an interdisciplinary approach drawing on fields such as artificial philosophy, biology, the cognitive sciences, artificial intelligence and, more specifically, the NBIC technologies (explained below) in order to assess, discuss, and ultimately *communicate* one of the most burning scientific topics of our day and age: the likelihood that sometime in the future human beings can upload their minds (a process known as *mind uploading* – in short, MU). This discussion includes the assessment of its most likely technical instantiation, a process known as Whole Brain Emulation (WBE) and also of the likelihood of human beings having in the future to deal with the explosion of Artificial Intelligence (AI) in such a way as to bring about a *Superintelligence* (SI) that could eventually entail an existential threat to the human species as such.

This is one of the most burning topics in scientific discussion today and, to my mind, it is also one of the most important. And this because, as will become evident throughout the thesis, it deals with the possibility of the radical transformation of humankind and, ultimately, with its viability, through completely different means, in the future. *More bit than bio*, then, means that in the future our fundamental core could be preserved through a very different kind of substratum, one which is more informational (bit) rather than

biological (bio). This of course entails a lot of assumptions, some of them technical, the others properly philosophical. But the point to emphasize here, for the time being, is the importance of the discussion: at stake is not mainly the future of the planet, as is the case with many scientific debates due to the current ecological threats, but the future of human beings as such, if and when faced with an intelligence greater than ours.

At the same time, and even if the body of discussion on this topic is evergrowing, the truth of the matter is that in Portugal, i.e., in the context in which this thesis is put forward, it is a topic almost completely absent from scientific and / or public debate. As such, this thesis aims to fill this gap in the current context of Portuguese communication of science.

In order to do so, this thesis uses three main instruments. The first, and more broad of them, is an overall effort to put forward a carefully argued critical assessment of the topics on which this thesis revolves. This entailed reviewing and discussing the most important literature in these topics in the several scientific fields that it touches on. The second, as I will explain in more detail further in this introduction, is the use of thought experiments for exploration purposes – these being a very important epistemic tool that has been used in the history of science and philosophy, and which allow us to imagine, to envision a given situation and to draw inferences and conclusions from that imagined – but not impossible – case. As for the third, and which involves a more carefully devised empirical methodology specific of the social sciences, it undertook a series of interviews with experts on these fields, carefully chosen to provide their specialized insight into these matters. These were the occasion of vivid dialogues – which can be read at the end of this thesis – and were a source of very important material for this research.

Readers should also approach this thesis with a caveat in mind: being exploratory in nature and dealing with very important topics whose evidence changes fast as technologies evolve, I see this work as only the beginning, and not the endpoint, of my research in this matter. My goal is also to open new avenues for the communication and discussion of this topic, first in the Portuguese context, but also, in the long run, in the overall discussion of these topics altogether. For this, the thought experiments I devised, and the interviews conducted strike me as important contributions to the field, and it is likely that different deliverables from this work – namely publications – will appear in the future. But now a more detailed introduction follows.

Nanotechnology, biotechnology, information technology and the cognitive sciences (NBIC) have in recent years been shown to enable human enhancements. This, in turn has raised a number of issues due to social and cultural implications, including the very definition of what it means to be human. Their impact has triggered a hype in movements, theories and beliefs linked to posthuman hypotheses such as transhumanism. Holding strong to scientific and technological trends related to human enhancement (and in particular enhancement of the brain), transhumanists arguably believe them to be leading humankind toward a new golden age where humankind could eventually live an eternal life in the form of a Prometheus of some kind. On the other hand, bioconservatism fiercely opposes human enhancement in the name of human nature* and instead regards the latter archetype more as a Frankenstein. Moreover, critics of transhumanism in general, envision these enhancements as dangerous possibilities leading to disruptive outcomes in human society or even as posing an existential risk.

Furthermore, humankind is arguably approaching a technological singularity in the next few decades, based on the acceleration of scientific and technological progress (Table 01), a point which stands for an impact so deep, mainly due to artificial intelligence and nanotechnology, that human life will be irreversibly transformed. Its outcome is at present also impossible to predict. This accelerated progress characteristic of human societies' recent history, carries profound political, economic and cultural changes and what is known as 'the Event' is generally linked to exponential growth in various technologies (with Moore's Law being a prominent example) as a basis for predicting that the singularity is likely to happen sometime still this century around the year 2045 (Kurzweil, 2005).

Table 01 - Shows the accelerated progress in generations. (Roco and Bainbridge 2006)

Generations	Several Key Advancements (human kind, tools and technology, communication)
-m	Cell, body and brain development
- 100,000	Old Stone Age (Paleolithic), Homo Erectus, speech
-10,000	Homo Sapiens, making tools
-500	Mesolithic, creating art
-400	Neolithic, agricultural products, writing, libraries
-40	Universities
-24	Printing
-16	Renaissance in S&T, accurate clocks
-10	Industrial revolution
-5	Telephone
-4	Radio
-3	TV
-2	Computers
-1	Microbiology, Internet
0	Reaching at the building blocks of matter (nanoscience) Biotechnology products Global connection via Internet; GPS/sensors for navigation
$\frac{1}{2}$	Unifying science and converging technologies from the nanoscale Nanotechnology products Improving human performance advancements Global education and information infrastructure
1	Converging technology products for improving human physical and mental performance (new products and services, brain connectivity, sensory abilities, etc.) Societal and business reorganization
n	Evolution transcending human cell, body, and brain?

In order to discuss human enhancement and posthumanity it is necessary to understand these NBIC enhancement technologies, which can be internal or external, temporary or permanent, cognitive or physical, individual or collective (Roco and Bainbridge 2006). Medicaments are a temporary individual internal enhancement whereas new organs or genes account for a permanent improvement although boundaries are eroding within technoscience making them hard to distinguish, especially if one recalls that technoscience is the theatre of an ongoing 'technologisation' or 'biologisation' of technology (Bensaude-Vincent 2009; Haraway 1991). Whether or not these technological developments are to engender Homo sapiens 2.0 as the next (man-made) evolutionary step or lead to a posthuman future (Fukuyama 2003), converging

technologies, along with the acceleration of exponential change (Kurzweil 2013) appear to challenge what it means to be human. One aspect of this larger-than-life paradigmatic shift that consists of the human challenging its own genetics, is the possibility that humanity might edit its own genetic code. The idea of living beyond genetics' predetermined life span and of existing outside biology altogether gains momentum with envisioned hypotheses such as mind-uploading (i.e., whole brain emulation or substrate-independent mind) of transferring or copying the brain's neurological functions onto computational code. Other options suggest augmenting the human brain with artificial general intelligence. (Koene, 2012).

Table 02 – Terminology adopted by Max Tegmark in his book *Life 3.0, Being Human in the Artificial Intelligence Age* used for definitions that are sometimes hard to describe and many times misused and misapplied. (Source: Tegmark, M., 2017).

Terminology Cheat Sheet	
Life	Process that can retain its complexity and replicate
Life 1.0	Life that evolves its hardware and software (biological stage)
Life 2.0	Life that evolves its hardware but designs much of its software (cultural stage)
Life 3.0	Life that designs its hardware and software (technological stage)
Intelligence	Ability to accomplish complex goals
Artificial Intelligence (AI)	Non-biological intelligence
Narrow intelligence	Ability to accomplish a narrow set of goals, e.g., play chess or drive a car
General intelligence	Ability to accomplish virtually any goal, including learning
Universal intelligence	Ability to acquire general intelligence given access to data and resources
[Human-level] Artificial General Intelligence (AGI)	Ability to accomplish any cognitive task at least as well as humans
Human-level AI	AGI
Strong AI	AGI
Superintelligence	General intelligence far beyond human level
Civilization	Interacting group of intelligent life forms
Consciousness	Subjective experience
Qualia	Individual instances of subjective experience
Ethics	Principles that govern how we should behave

Teleology	Explanation of things in terms of their goals or purposes rather than their causes
Goal-oriented behavior	Behavior more easily explained via its effect than via its cause
Having a goal	Exhibiting goal-oriented behavior
Having purpose	Serving goals of one's own or of another entity
Friendly AI	Superintelligence whose goals are aligned with ours
Cyborg	Human-machine hybrid
Intelligence explosion	Recursive self-improvement rapidly leading to superintelligence
Singularity	Intelligence explosion
Universe	The region of space from which light has had time to reach us during the 13.8 billion years since our Big Bang

This *Terminology Cheat Sheet* presented above, taken from Max Tegmark's book *Life 3.0 (Being Human in the Age of Artificial Intelligence)*, elegantly sums up the language, the categories and levels of the main subjects explored throughout the thesis. (Tegmark, 2017).

Taking into account the context already provided, allow me to move on to briefly describe the content and structure of this thesis. It comprises five different parts, of which the first four are thematic chapters, exploring Thought Experiments (I), Whole Brain Emulation (II), Mind Uploading (III), and Superintelligence (IV), while the latter (V) consists of the analysis of interviews done with experts in this area. Parts II to IV are of course thematically linked, insofar as WBE, MU and SI are closely connected hypotheses which are here explored *in tandem*. The reason why the first part is dedicated to thought experiments relies on the need to present and justify one of the main ways in which, throughout the thesis, I explore these topics, which is precisely by means of devising a thought experiment. A brief explanation of the main objectives of the thesis and of the methodology used in each of the parts of the thesis now follows.

General Objectives of the thesis

These are the general objectives that I propose to achieve in this thesis:

- To explore the link between the convergence of NBIC technologies and the emergence of human enhancements such as mind uploading and the creation / development of superintelligence.
- To design Thought Experiments that are applicable to the main themes of this thesis, i.e., to help confirm or reject some premises regarding mind uploading and superintelligence.
- To scrutinize the state-of-the-art of Whole Brain Emulation and thus test its feasibility and estimate its timeline.
- To investigate the outcome and impact raised by Mind Uploading regarding consciousness and personal identity issues, eventually altering *what it means to be human*.
- To critically analyze both the claims of Mind Uploading's proponents and of its skeptics so as to cover it as extensively as possible and thus enable an *open debate over open questions*.
- To apply the MU and SI thought experiments described in Part I, by analyzing their content in the context of *mind uploading* and *superintelligence*, and therefore verify their internal consistency, hence validation.
- To establish reasonable inferences exposing *flaws and missing links* hindering the course toward Mind Uploading's feasibility in order to shed some light leading for possible alternatives in that path.
- To clarify Mind Uploading's purpose and meaning, thus trying to prove its importance and significance for humankind's evolution.
- To discuss a possible clash between an Artificial Superintelligence and our own intelligence and argue in favour of Mind Uploading as a technology to help humankind in that competition using the designed thought experiments to help envision what is at stake in that dispute.
- Finally, and this is the broadest objective and the one to which all the above-mentioned objectives concur, this thesis aims to contribute to the dissemination

of these scientific themes, through a carefully argued discussion of the (technical) feasibility and (ethical) desirability of these possibilities.

Methodology used in the different Parts of the thesis

Part I

- The methodology of Part I (*thought experiments*) consists of a narrative way to explore, by resorting to productive imagination, some conceptual possibilities whose implications can be easily apprehended in narrative fashion. And these possibilities, in turn, can be explored, in an argumentative way and help sustain the main claims of this thesis. This is precisely what is done afterwards, insofar as the scenarios of the thought experiment devised then are mentioned in the following parts of the thesis, in order to illuminate what is at stake when discussing given topics and envisioning future possibilities. Furthermore, thought experiments are important instruments in the teaching and communication of science (for instance in the domains of biology and the neurosciences) insofar as, taking place “virtually” and not needing laboratorial testing, they help to transform people’s representations (Calafate, 2019). Thus they have a direct impact on the beliefs concerning possibilities that cannot be empirically verified for the time being but that might be true and apply in the future. It must be mentioned, however, that before the different scenarios of the thought experiment are presented, this part introduces to the topic of thought experiments in general, demonstrating their importance in the history of science and philosophy alike, and showing how they function.

Parts II, III & IV

- The methodology used in this thesis aims to follow the kind of *SALSA* (*Search, Appraisal, Synthesis and Analysis*) framework in order to form an *analytical critique*:

A *literature review* is done, evaluating and synthesising fields of research and highlighting what has been done, what is generally accepted, what is emerging and what

is the current state of thinking on each topic, to identify gaps and argue for the need for further research using a comprehensive approach of the sources;

The literature review also entails a *critical review* evaluating specific and single sources based upon conceptual, theoretical and methodological considerations in order to go beyond *mere description* and to include some degree of analysis and conceptual innovation and also to contribute to an open debate of these themes;

Part V

- The methodology used here consists of content analysis of *semi-structured interviews* in order to contribute to SC (Science Culture) and PAS (Public Awareness of Science) thus contributing to science culture among the general public as well. This fulfills a basic role in the communication of science, that of outreach. As will be explained in more detail below, the interviews were conducted with several experts in the different fields relevant for the topics discussed in this thesis, with the sample being constituted by rational choice. These interviews and the analysis of their content can be seen as examples of meaningful science communication and outreach insofar as they bring together several key experts and engage them in a discussion of several important aspects, including my own thought experiments.

For a fuller account of the steps undertaken when devising this methodology and its justification please see the content of each part which, in a nutshell, is as follows:

Part I _ Thought Experiments

The first part of this thesis presents thought experiments (*Gedankenexperimente*) regarding superintelligence and mind-uploading as a true possibility in the near future. The research mentioned in this thesis investigates the use of innovative technologies and design-based approaches (e.g., whole brain emulation, artificial general intelligence, molecular nanotechnology), and inquires into their potential for (re)defining humankind.

Since many of these subjects (mind uploading & superintelligence) do have a highly speculative weight, thought experiments as sustained arguments that can entail epistemic value, exhibit a valuable insight for exploring reasonable outcomes, discuss and communicate fundamental scientific ideas. Their applicability, i.e., the purpose and target of their *rationales* are explored throughout the various chapters, namely in the mind uploading and the superintelligence parts aiming to sort out some answers to the proposed issues. If well designed enough they might confirm or even challenge some of the claims entailed in these hypotheses. At the least they should be able to provide material to extrapolate beyond established facts and help argue in favour or against some of the topics approached in this dissertation.

Before presenting these thought experiments, the current state-of-the-art is described as an introduction to these subjects and methodology. Some of these thought experiments have been presented as *work-in projects* to the interviewees for the semi-structured interviews made for this thesis in order to obtain a feedback to help devise persuasive and cogent arguments. As explained above, they aim to support the objectives of this dissertation by exploring and communicating about superintelligence and/or mind uploading as core topics of the thesis.

Ethical Disclaimer

As an ethical disclaimer it has to be mentioned that these TEs might in some scenarios present outcomes that are ethically problematic, as they might go against the grain of the ethical intuitions of some people. This is mainly due to the speculative nature of these TEs and therefore also to keep their possibilities as *open questions* for it cannot yet be anticipated the extent to which they would actually push ethical borders in practical terms. The ethical implications of the subjects discussed have of course been taken into account on several occasions throughout the dissertation (and notably in the [section 6.2](#)) yet surely not as much as they eventually should be addressed when approaching subjects as controversial as the ones explored in this thesis. However, it was a deliberate intention to focus on the scientific, technical and sometimes philosophical aspect of these subjects in a perspective of what can be done, what might happen and what is doable or not. If I were to delve in a deeper ethical approach than the one I took, I would probably

end up writing another thesis. This being said and as an argument in favour of those speculative ventures, I would like to recall an historic example:

In the late seventies, Smallpox menaced humankind inasmuch as its pandemic outcome threatened to be devastating in terms of world population with a death rate of about 30% and estimated to be 3-5 times more infectious than the COVID-19 corona virus of today. By 1980 the WHO managed to eradicate Variola successfully through massive vaccination with a quite *risky but incisive approach*. By then, vaccines were not developed with the safe and thorough guidelines they are today, and it is arguable how devastating it could have been if things had gone wrong. The risk had to be taken in the name of humankind.

This is not to advocate that anything that can be done should be done or, really, any kind of principle other than the fact that, as history proves, when faced with catastrophic possibilities, some difficult moral decisions (as the famous trolley thought experiment exposes) have to be made and a weighed balance of a kind of risk/benefit assessment might be justifiable under the pretext of human survival. What is at stake in this discussion will become clearer after the Thought Experiments are presented and discussed throughout this thesis.

Part II _ Whole Brain Emulation

The second part of the thesis revolves a lot on the feasibility of achieving a whole brain emulation and puts forward a thorough investigation of the tools involved in WBE and discusses its feasibility in the spectrum of an up to date roadmap towards that target. It identifies the convergence of the nano-bio-info-cogno fields (NBIC) as the appropriate domain to focus on the scientific and technological developments concerned with human enhancements such as mind uploading and the development of an eventual augmented intelligence (superintelligence), and the creation of SIMs (Substrate Independent Minds) in the following chapters. Its clarification provides ground to analyze ethical, social and political outcomes linked to it.

Part III _ Mind Uploading

In the third part of the thesis, mind uploading is analyzed at various levels that are ultimately changing our very concepts regarding consciousness and personal identity intimately linked to that eventual achievement. Hence, consciousness and personal identity are thoroughly explored and discussed in this chapter and although these two are still (*and always*) an ongoing controversial scientific and philosophical challenge, some clarification might come to them precisely through the development of mind uploading as a technology / technique. It will be argued that even in the event mind uploading would not become feasible, or at least not as fully functional as envisaged today, its impact as a concept and as a work in progress is worthwhile discussing.

Part IV _ Superintelligence

In the fourth part of the thesis, a technological singularity and its logical download of an augmented intelligence in an incommensurable magnitude are discussed as a strategy to address questions and controversies that have risen through enhancements toward a posthuman condition. With a superintelligence, mind uploading techniques could become real and alter our own selfhood in such a disruptive way for our anthropological self-understanding that would make it justified to ask, in an analytical and critical fashion, what it means to be human. Promise or peril, utopian versus dystopian future, whichever should come true, posthuman discussions have become a main topic on today's agenda regarding the societal, political, and ethical implications of these NBIC technologies.

Vernor Vinge developed the concept of the technological singularity in his seminal essay "The Coming Technological Singularity" (1993). What types of creative issues might there be for early posthumans? Would the intellectual and creative pursuits of past geniuses be immeasurably and incomprehensibly surpassed by a posthuman creativity augmentation? One in which the idea of creativity as we know it today would become quite different and expand exponentially in ways not yet imaginable. This expansion could be one possible result of the human brain being augmented with AGI and connected seamlessly with other minds, thus forming a type of joint of interconnected creativity mosaic among two or more minds.

Part V _ Semi-Structured Interviews – Analysis

In the fifth part, interviews with leading experts (mainly neuroscientists, computer scientists and philosophers) were held regarding the problems and questions exposed and discussed throughout the thesis and the impact on humankind and *on what it means to be human*. The previously selected variety of experts (sampled by rational choice) was intentionally devised to reflect a degree of societal knowledge regarding humankind's possible future outcomes. The interviews outlined with predetermined questions with two different possible outcomes have been designed to be objectively linked to the topics and issues investigated throughout the thesis. The presentation of the thought experiments was important to create a field of discussion concerning the implications of mind uploading. The data obtained from these interviews are herein qualitatively analyzed and add an insightful knowledge to the main theses discussed throughout the thesis. The transcriptions of the interviews can be found in the Appendix.

Final Notes

Printed books extended our memory and allowed us to externalize our thoughts. These outsourcing capabilities are now exponentially increased through computers and through the largest knowledge base ever made by humans: the internet. Humankind is effectively outsourcing memory and ultimately even outsourcing intelligence. (Clark and Chalmers, 1998). AI systems are developing so dynamically it has become difficult to update their progress. AI will inevitably outperform humans in every task within the next couple of decades and automate all human jobs in another couple more. (Grace et al, 2017). Microsoft announced an investment of over 1 Billion \$ in 2019 to develop AI together with OpenAI. According to OpenAI's CEO, Sam Altman, the creation of AI could be the most pivotal event in human history: "Those that are genuinely concerned about AGI and ASI as a potentially uncontrollable threat to the survival of the human species argue that ASI is not a disruptive technology like the printing press and electricity, but that it is biologically significant because it is concerned with intelligence." (Pohl, 2015, p. 11). It was precisely *our superior intelligence* that carved humankind's path until today and now the possibility of machine intelligence, orders of magnitude higher than ours is about to take that lead and it's starting to divide our society as shown on Figure 01. It is therefore worthwhile to anticipate an eventual singularity, even if, as many authors argue, its

probabilistic chances would be as low as 1% - its associated risks are very much worthwhile paying attention to. (Bostrom, 2014; Goertzel, 2015). These are the topics we are now going to delve into.

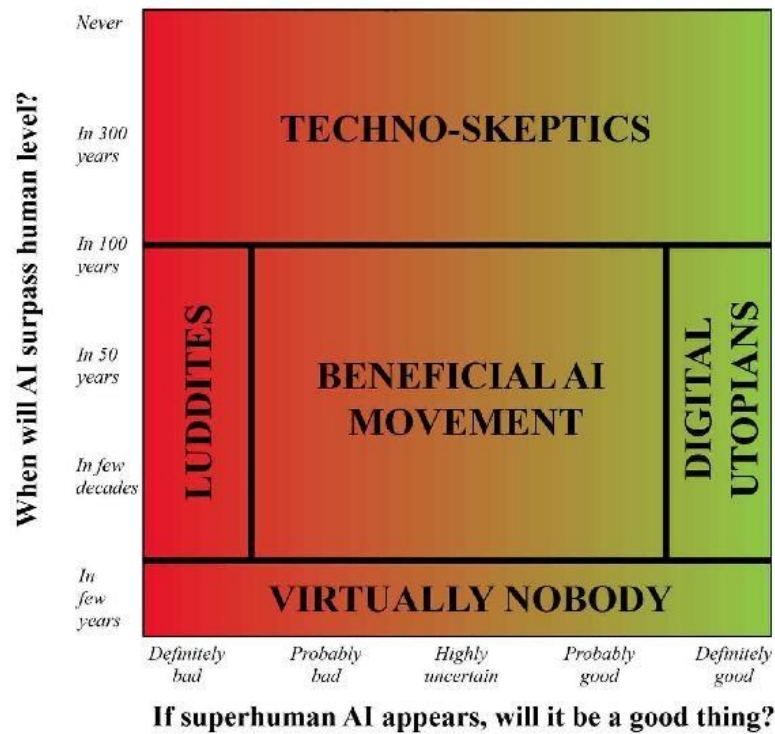


Figure 01: Concerning artificial intelligence, two of the main questions asked are: When will AGI override humankind and will it be beneficial to us? (Source: Max Tegmark, 2017).

* The term *human nature* is used as the concept that denotes the fundamental dispositions and characteristics — including ways of thinking, feeling, and acting — that humans are said to have naturally and is often used to denote the essence of humankind, or what it 'means' to be human. This is a general definition broadly used and accepted *by default* and throughout this thesis I haven't dared to redefine it or even explore its philosophical connotations on purpose.

PART I _ MU+SI Thought Experiments



I took a test in Existentialism. I left all the answers blank and got 100.

— Woody Allen

1 _ Thought Experiments

1.1 _ Intro to thought experiments

Although only recently introduced as a term in contemporary times in the 19th century, thought experiments have been around for a long time with historical records showing us that already in the philosophies of Plato and Aristotle something akin to thought experiments already shows up. But it was not until the eighties of the 20th century (in fact the first conference on thought experiments only took place in 1986) that they got philosophical attention and since then, as even now, their fundamentals as to what they are and what for, are still very much in debate.

Nevertheless, it is generally well accepted that Galileo introduced them as a tool with evidential significance as scientific experiments. In fact, Kuhn views thought experiments as essential analytic tools that in critical scientific times help to trigger off scientific revolutions, and in this way promote basic conceptual reforms in science. “From thought experiments most people learn about their concepts and the world together” (Kuhn, 1977, p253). So, in his view they were not used to generate facts but to promote breakthroughs between scientific paradigms in critical times of science. Furthermore, Kuhn’s paradox that thought experiments frequently have novel empirical import even though entirely conducted in thoughts, opened up a quite enduring debate (still going on today), mostly carried through by James Robert Brown and John D. Norton. Without dwelling too much on this topic, because it is beyond the scope of this thesis, we can sum it up by mentioning that Norton and Brown actually agreed with Kuhn that thought experiments do play a major role in scientific revolutions and that their main source of information definitely relies on scientific record. Their debate only started off (and triggered Kuhn’s paradox) regarding the nature of thought experiments, because Brown believes them to be actual thought experiments that provide direct access to truths about laws of nature whereas Norton claims them to be just arguments.

This debate revolving around the epistemic reach of thought experiments went on for a couple of decades with other actors supporting each of the contender’s views. Without entering myself in the debate I will adopt some definitions concerning thought experiments herein described as being generally accepted thus in accordance with my

own ensuing thought experiments (3a - iterating human predictor response / 3b - iterating human intelligence / 3c - delusional reality), also herein presented in this chapter.

Throughout history thought experiments have been particularly successful in philosophy, and in science mainly in physics. In this case, and given the scientific background of the thought experiments I present (3a, 3b & 3c), as well as my own scientific background, the aforementioned thought experiments are scientific in nature (rather than merely imagined such as it sometimes happens in philosophical thought experiments) and thus designed to mainly reach the scientific community. Having said this, it must also be admitted that in spite of having some scientific, objective grounding (in that, to our knowledge, nothing prevents these experiments from being real possibilities) these thought experiments have as their starting point some assumptions that might be philosophically controversial, as well as ethical issues that need to be open to discussion, and I will come back to this issue later in this thesis.

So, in the case of science, a thought experiment is constructed to produce an outcome that pretends to have scientific import, and this differentiates it from fictional imagination. This outcome, in turn, and just as it happens with empirical experimental outcomes, will be subjected to interpretation and possibly even to further investigation. It often highlights a problem. But this does not mean it necessarily provides a solution by itself; rather, it often pushes for additional work to solve the problem. It may not even be its goal to provide solutions. It may serve the purpose “to introduce a conceptual distinction, to refuting a usually admitted thesis, to testing a declaration by highlighting a paradox, or even proving a previously unknown result” (Sorensen 1992, p135–52).

The ontological question of what thought experiments are is entangled with the epistemological question of understanding where the knowledge that they bring may come from and what kind of knowledge exactly they can bring. Thought experiments explore the outcomes of counterfactual scenarios in a digital context. (Roux, 2011).

These counterfactual scenarios display a kind of opacity on which the thought experiment operates, which is argued as one of the reasons for not being reduced to simple arguments. Thought experiments do not have to take place in reality as opposed to real empirical experiments. If the thought experiment does not present much, or any counterfactuality at all, this means that it nears reality inasmuch that it could take place in physical reality but for some reason is not (or not yet). On the other hand, if it deals

with a lot of unreal scenarios, that for some reason cannot be carried out whether for technical motives or due to some kind of inaccessibility to perform them. In this case, they entail a strong counterfactuality. Jean-Yves Goffi and Sophie Roux use the EPR thought experiment to support this argument. The Einstein-Podolsky-Rosenbaum *Gedankenexperiment* (EPR) dates back to 1935, in a time in which its proposals about the completeness of quantum mechanics could not have been carried out physically. Another example is Stephen J. Gould's *Wonderful Life* (1989) thought experiment of notion of rewinding and replaying time to find out if evolution would have been the same as it is. It stands for an even stronger case of counterfactuality for it not only depends on technical means but also on an almost impossible kind of intervention over time itself, which nevertheless does not make it a lesser thought experiment.

Besides presenting an imagined scenario, and this being counterfactual, a thought experiment also presents another typical characteristic, that of being determined in a specific cognitive context that ensue the role of an argument. Unlike logical arguments, thought experiments involve scenarios in which non-propositional knowledge intervenes, whether it is called implicit knowledge, tacit knowledge or background knowledge. (Gendler, 1996; Gendler 1998, p404–8). It is rather a process that supplies premises for an argument or that consists in unfolding an initial scenario. (Sorensen, 1992; Häggqvist, 2009).

Thought experiments are imaginative exercises designed to disclose what happens when certain conditions are met in a realm in which ideas are bound by constraints previously set. Hypotheses can then be contrived, elaborated and tested in a laboratory called the mind. These constraints, to which thought experiments are subjected, even if tacit, are recognized, discussed, reviewed and eventually validated by the communities to which they are addressed whether they are scientific or philosophical. In designing a standard experiment, a scientist may begin by performing a thought experiment running through the expected course of events in his or her head before attempting to implement it in the lab. The difference lies in the fact that thought experiments sharply truncate the experimental process, leaving out the implementation step. Sometimes the experiment envisioned cannot be carried out whether for physical, ethical or any other reasons making it impossible to run through just as happens with regular empirical experiments.

By imagining the experience of a person riding in an elevator in the absence of a gravitational field and at rest in the presence of a gravitational field, Einstein showed the

equivalence of gravitational and inertial mass. To actually execute the experiment would be physically unfeasible. Still, sometimes, the infeasibility of an experiment translates into the infeasibility of finding out a particular fact. Einstein's thought experiment is effective by challenge: Suppose the specified conditions were met. How could a subject tell the difference and know in which situation he was? We can have the so-called evidence of the equivalence whenever we cannot, for scientifically principled reasons identify a way to tell the difference. Hence, by failing in detecting a difference we can assume our theories to be close to correct. Sometimes the imaginative rehearsal dismisses the necessity of actually carrying out the experiment.

Without physical implementation, Galileo's thought experiment discredited the Aristotelian claim that the rate at which bodies fall is proportional to their weight. By exemplifying the inconsistency, Galileo's thought experiment proves the Aristotelian account to be wrong. An empirical experiment would not have made Galileo's case better though. Potential sources of interference such as distance and duration of the fall and the presence or absence of air resistance amongst others are omitted thus blocking these challenges and demonstrating Aristotle's inconsistency regardless of the conditions under which the experiment was conducted.

Sometimes “the thought experiment is preferable to an actual experiment because it is invulnerable to a host of potentially misleading challenges that an actual experiment would face. Even in the empirical sciences, not every question can or need be answered by direct appeal to observational evidence” (Catherine Z. Elgin, 2017, p232).

This underlines the role of thought experiments in the discussion and communication of science, which is one of the main goals of this thesis. If well drawn, a thought experiment is a catchy idea that sticks to one's mind. People will more easily recall such an imagined example than a fully developed theory. For instance, perhaps not everyone will be acquainted with the intricacies of utilitarianism; but everyone will understand what is at stake in the trolley problema. And this is one of the reasons why using such experiments can be a successful strategy in the communication of science, and also one of the reasons why they are used in the context of this work.

1.2 _ The relation between thought experiments and empirical experiments

As mentioned above, in science, thought experiments are developed to provide novel insight into real-world phenomena or to confirm or reject a hypothesis. But just as discussed earlier, Kuhn's paradox maintains, although various attempts have been presented in order to explain how an experiment performed only in the mind, and in the absence of new data, can actually find application in the real world.

One of those attempts made by Nancy J. Nersessian et al, concludes that thought experimenting is a form of mental modeling, a form of simulative model-based reasoning, where inferences are made through constructing and manipulating models, whether of conceptual, physical or computational nature (Chandrasekharan and Nersessian, 2015). Their hypothesis states that a successful thought experiment can reveal, highlight or disclose empirical data previously hidden. Furthermore, they claim that mental modeling has over time in our evolution proven to be fundamental in human reasoning when dealing with problem-solving tasks and that this feature stands for an efficient means of anticipation, which also has a survival significance. The inference one can take out of a thought experiment applies to other situations of the same kind just as happens in science with empirical experiments. Generally, they isolate and extract what is common to phenomena of the kind under investigation and not just that one particular phenomenon.

Scientific empirical experiments do not actually have an ultimate truth-seeking goal. They enable new interpretations, inferences and hypotheses to quantify the data collected, to mathematize it. Catherine Elgin argues that "science only needs to meet its cognitive goals, it does not need to produce truths for their epistemically outcomes to be significant" (Elgin, 2004). Thought experiments enable science by reasoning a feasible simulation of some representative situation which is the result of mental modeling of that same representation derived from real-world experiences and of our conceptualizations of them. (Nersessian, 2018). However, since an empirical experiment does get carried out in physical reality, it never gets confused with an argument as may be the case of thought experiments. The experiment itself does have an argument too, but the scenario in which it takes place is real as opposed to the thought experiment's scenario that actually unfolds at the same level as the argument does, hence blurring its concepts.

Experiments are characterized by the manipulation of independent variables and the use of controls. Empirical experiments are suitable in some cases whereas thought experiments can be more adequate in others. Thought experiments can also help determine what kind of empirical experiments to perform and how to interpret them. Good empirical experiments are consistently reproducible because they isolate from and control for outside influences. They start with at least one scientific hypothesis that is falsifiable and offers predictions that can be experimentally tested and whose accuracy may end up determining whether the hypotheses themselves are supported or rejected by the data. Likewise, thought experiments can isolate specific variables to explore the relationship between them and create a hypothetical situation in which all the parameters are set in order to limit consideration to the specific decisions or phenomena being studied and in this manner lead to a general statement of principle thus helping understand how something would be in an ideal case. Like predictions, these statements of principle can then be tested with further experiments to find circumstances in which they do not hold or instead to show precisely that they do so.

These two types of experiment complement each other well because the things thought experiments are good for, i.e. making falsifiable predictions and assessing the implications of facts and principles on real-world circumstances, are the ones empirical experiments do poorly. There is a tradeoff between internal and external validity that researchers in empirical experiments are confronted with, which in turn also makes them strongly dependent on a theoretical conceptualization to better evaluate their internal consistency. Similarly, empirical research provides a way for philosophers to determine if the assumptions on which they construct their theories are consistent with common knowledge. Thus, it is understandable that many of the greatest thinkers were both philosophers and scientists. Many of humankind's greatest achievements have come from people who were willing and able to harness the analytical and predictive power of both disciplines and use them in tandem. It is of the utmost importance for scientists to have an analytical thinking capacity tied to a philosophical knowledge. Conversely, also philosophers with a scientific and empirical research interest, have an appreciable edge.

Besides empirical and thought experiments, simulations too, are a kind of experiment. As happens with thought experiments, simulations also take place in hypothetical worlds, without intervening in the natural world as opposed to empirical experiments. Johannes Lenhard argues that thought experiments are epistemically transparent and use intuition in their cognitive process whereas simulations relate to automated iterations of formal

algorithms making them epistemically opaque experiments. He argues that this difference displays epistemological ramifications. The thought experiment can fail not because of an unwanted outcome but because if its outcome is unclear, if its outcome does not make sense given its premises, or if it is not a logical outcome, then it simply will not allow a clear judgement and therefore it will not work as a thought experiment per se. This is why it has to be epistemically transparent whereas for some authors (Di Paolo, Noble and Bullock, 2000), simulation experiments could even be held as opaque thought experiments by delivering hypothetical assumptions in a computational way. While it is true that simulation experiments are similar since both are a sort of model-based reasoning, it is however also true that unlike thought experiments, simulations are not tied to intuition although being able to use a range of algorithmic procedures. Therefore, Chandrasekharan and colleagues conclude that simulations will likely outpace (and eventually replace) thought experiments. (J. Lenhard, 2018).

Epistemically, however, there seems to be an important difference setting these two kinds of experiments apart. Simulations use iterations under controlled variations to exhaust all possibilities leading to a sort of compendium, the so-called atlas iterations whereas thought experiments are tied to the so-called convergence iterations. Lenhard argues that when performing a thought experiment, its initial intransparency will ultimately be eliminated by the repetition of its execution which does not occur in simulations since their iterations do not eliminate opacity but rather explore the space of possible model behavior. (Lenhard, 2018). This, of course, does not tarnish the value out of the simulation experiment since computing models can automate iterations in such a way as to compensate for the lack of transparency gathering results that can substitute intuitive certainty. Iterative algorithms utilize computational power and can work where thought experiments cannot. “Processing in intuition requires transparency, which computer-automated iterations often cannot provide. However, the insights simulations gain from collecting many results are something thought experiments do not provide”. (Lenhard, 2018, p11).

From the above it would only be fair to claim that thought experiments and simulation experiments complement each other, as has also been the case with empirical experiments. This State of affairs would actually perfectly fit the thought experiments presented here and that consist of (3a) iterating human predictor response, (3b) iterating homo sapiens intelligence, in the sense that it entails a simulations component that

nevertheless has to be intuitively interpreted regarding its outlined scenarios specific context.

1.3 _ Some classical thought experiments

In the two major fields contributing to / with thought experiments (i.e., philosophy & science), we can find many examples that became famous for one reason or another. Some of them are described below because they directly relate to issues concerning this thesis, or because they resemble to some extent the thought experiments outlined for this thesis.

In philosophy, there is Searle's' Chinese room, Putnam's brain in a vat, Nozick's experience machine, Rawls' original position, Jackson/Dennett's color scientist, Chalmers's zombies, and Foot's trolley problem, to name just a few.

In science, we can find Lucretius' throwing a spear at the edge of the universe, Maxwell's demon, Einstein's elevator (and train), Schrödinger's cat, Newton's bucket (and cannonball), Heisenberg's microscope and Galileo's falling bodies.

Galileo's thought experiment showed an inconsistency in Aristotle's theory but also that any theory that took the rate at which bodies fall was proportional to their weight would be inconsistent. From this it follows that the rate at which bodies fall is independent of their weight hence a thought experiment not only discusses a theoretical or conceptual allegiance but can also exhibit a feature of the world that is a consequence of a modal fact. In this case, the fact that the rate at which objects falls is independent of weight. The objects could never have fallen in any other way. Thought experiments afford epistemic access at least to theoretical, conceptual, and modal matters. By fixing certain parameters they provide a description of the experimental situation that sets out all and only the features considered relevant and work out the consequences. (Catherine Z. Elgin, 2014).

Einstein imagined he was riding on a beam of light looking at another beam parallel to his own. Classical mechanics would tell him that the second beam should appear to be still, yet according to electromagnetism light's speed must always be $3 \times 10^8 \text{m/s}$. Both classical mechanics and electromagnetism claimed to be universal, meaning that the laws were the same for all observers, but Einstein's thought experiment led him to

conclude that either classical mechanics or electromagnetism was wrong, or that neither was universal and hence special relativity was born. Since empirical observation itself is relative, undergoing a scientific experimentation even if humanly possible would not have made a difference, hence the need to recur to a thought experiment. That was Einstein's sole available tool to come up with what would prove later on to be one of the most important theories in humankind. Without philosophy, the theory of relativity and the scientific breakthroughs that came with it would have been impossible.

It does seem appropriate to include in this section a review of some of the above mentioned famous thought experiments whether because of their possible similitude to my own thought experiments herein presented, or even because some of them address some of the controversial issues explored in this thesis involving scenarios such as identity, intelligence and consciousness.

Brain in a Vat is another classic famous thought experiment that finds its origin in Descartes' evil demon problem, whereas its update is credited to Putnam. The TE postulates that your brain has been placed in a vat connected to a computer which actually tricks you into thinking of yourself and everything around as you usually do. In Descartes' TE, the computer was actually some kind of demon feeding false information to your senses. So, whether Descartes' Demon, or Putnam's computer, you force yourself to ask the same question: Are you a brain in a vat? Are you being fooled in your reality?

The paradox of this TE is that you cannot actually determine for sure whether you are in fact a brain in a vat or in your skull, as you surely think you are. It could be an entire fictitious world that, to the captive brain, would feel completely normal, thus one would never know if experiences were real or an illusion. Putnam side-steps the problem of skepticism. He says that one should count the things one believes to experience as being knowledge, which could still mean one could be a brain in a vat. For instance, if one thinks to have eyes although being a brain in a vat, it still is true, but the eyes one's referring to are virtual. This would be like considering our own proposition to be a true claim based on what is epistemically available. Actually, criticism of Putnam's answer to skepticism relies on the fact that he ignores skepticism by merely assigning truth values to statements about things that may be outside of the epistemically available world. For instance, a brain in a vat could never truly claim to be a brain in a vat. This TE impacts on quite a few different fields of thought. Its ethical implications such as the fact that if

one would in fact be a brain in a vat, and nothing outside the world actually being true, then one could easily misuse everything around. It also carries implications to the meaning of life, and existence as is similarly discussed in the delusion of reality (TE3c/d) of this thesis. It questions what it means to be human. (Keith DeRose, 2000; Brueckner, 1992).

Searle's Chinese room is also a classic TE regarding philosophy of mind that argues that computers cannot have minds of their own, therefore are unable to generate consciousness. The TE tries to prove that computers could in fact be programmed to mimic human behavior (very much like in a Turing's test) but cannot really understand their own generated simulation: "The computer is not merely a tool in the study of the mind, rather the appropriately programmed computer really is a mind in the sense that computers given the right programs can be literally said to understand and have other cognitive states" (Searle, 1980, p. 349).

In his skepticism of the so-called strong AI, or what we now would define as AGI, Searle imagines a non-mandarin speaker inside a room, who is given a bundle of boxes containing Chinese writings with characters and symbols he cannot understand, as well as a set of instructions according to which the symbols may be formed. We are then to imagine a mandarin speaker outside the room that passes him messages so he can follow the instructions from the book to select an appropriate answer. Now, the person outside would be tricked into thinking he or she is having an actual conversation in mandarin with Searle, who as he states from the beginning doesn't of course know a word in Chinese. The TE suggests that however well you program a machine to follow the rules it still doesn't understand Chinese inasmuch as Searle doesn't either. It would only simulate to have that knowledge without actually even being aware of it. It can perform an output without really knowing the meaning of its input therefore it cannot be said to be actually thinking.

Now, this has of course some important implications to functionalism, which is an important issue regarding whole brain emulation and mind uploading, as will be addressed in this thesis.

Materialism unlike dualism argues that the mind and body are not distinct, rather the same, and mental states are just physical interactions. Logical behaviorism, on the other hand, states that all mental states can be explained in terms of behavioral concepts.

Identity theory argues that mental events are actually physical events and sees the mind as a physical entity where mental states are constituted of physical chemicals of the brain. Now, functionalist theories of the mind, unlike logical behaviorism, hold that mental states are internal states of thinking entities and not just behavior. Unlike the Identity theory, functionalism does not claim the mind and the brain are the same, although both are physical. For functionalists, the mind is to the brain what in the computer's software is to the hardware. Computer programs are not physical entities occupying space, although they have been created by such physical or chemical properties. So, in a parallel analogy, for functionalists, the mind is not a physical or chemical property even if it has resulted from such a property. In essence, mental states are numerous sensory inputs with behavioral outputs. If we think of computers as physical devices with electronic substrates that perform computations on inputs to give outputs, then the brain is a physical device that performs computations on inputs which produces behavioral outputs. In Searle's reasoning, this is where functionalism fails inasmuch as he claims that there is this missing part, which is understanding, hence the Chinese room TE. (Hauser, 1997, p199–226; Cordio, 2008).

The trolley problem is a version of an ethical dilemma devised by Foot in 1967. The scenario envisaged in this TE is that of a runaway trolley heading towards five workers who can't move out of the way. You happen to be standing next to a switch that will divert the trolley onto a second track. But that track has a worker on it, too, but just one. The TE explores our moral intuitions classically dividing right from wrong. It then has slightly different versions in order to complicate even more our dilemma. In one of such subsequent versions you actively and physically have to push a person to actually save the other five. This divides moral thought schools into utilitarians, who think that you have to sacrifice the one worker in order to save the other five on the other track, and deontologists who claim that you can't kill an innocent person even if it implies saving the other five. Research shows that generally people do accept you to take the utilitarian perspective and save the five workers, but in the second version of the trolley problem, people generally condemn you for actively pushing and killing the one worker, thus relating more to the deontological perspective. So, in the footbridge case, which is the version involving physical contact, harming one to save many is less acceptable than in the switch case version that does not involve physical contact. We then face a normative problem, which approaches the scenarios in which we should in fact approve such tradeoffs. "The longstanding hope is that a solution to the normative Trolley Problem will

reveal general moral principles. Such principles, in turn, may apply to challenging, real-world moral problems such as those encountered in the domains of bioethics (Foot 1978; Kamm 2001), war (McMahan 2009), and (most recently) the design and regulation of autonomous machines such as self-driving cars (Wallach and Allen 2008).” (Greene, 2016).

This thought experiment’s popularity derives from the fact that it forces us to think about how to choose when there really are no good choices at all. This situation is also addressed in this thesis’ TE 3c/d the delusion of reality where post-humans are confronted with the dilemma of remaining human but in an unreal kind of environment as opposed to giving up being human, in order to pursue that real existence.

Finally, it must be said that thought experiments are useful because they allow us to see possibilities to which we are not accustomed. Indeed, we are hardwired to think mechanically by memorizing stuff we learned since our early education and thus creating strong thinking patterns. But this pattern recognition makes it hard to be innovative and approach problems in a creative mode particularly when confronted with unusual data. These patterns once activated lead our thinking toward something that has worked similarly before and even in the absence of much of the information our brains automatically correct and complete the void by selecting a previous already applied and successful pattern. Thought experiments can deploy different thinking patterns and enable them to explore dissimilar and unrelated subjects and information.

1.4 _ Some rules regarding the conception of a thought experiment

Thought experiments can be designed to support or refute a scientific theory and most of them start from empirical background knowledge previously accepted: “To conduct a thought experiment is to make a judgment about what would be the case if the particular state of affairs described in some imaginary scenario were actual” (Gendler, 1998). Many times, and because of their nonempirical nature, they can be classified somewhere in between science and philosophy and also because of the type of questions they posit. In Kuhn’s account, thought experiments work by providing scientists a way of retrieving memories of things they have seen before but that only became evident precisely through them. These memories are uncovered by the so-called narrative structure of the

thought experiment. Although Norton considers them to be just dressed-up arguments as mentioned earlier, he does admit them to be a series of propositions without epistemic loss that lead to a conclusion via inferences of a recognized sort. So, the knowledge they entail and can convey, whether as thought experiments or as dressed-up arguments, as he calls them (Norton, 1996), does have that value hidden in the narrative. The narrative is of the utmost importance in conceiving a thought experiment as we'll see later on.

Thought experiments characteristically present us with a series of “what if” questions in order to conceive a coherent model of a scenario imagined according to the answers obtained. An internally consistent model can deliver several possible outcomes as a dynamic representation of the scenario and it is this consistency of the model that differentiates a thought experiment from fiction and enables the experimenter to answer consistently to the “what if” questions elicited. They can then reveal implicit consequences of theories by seeking what our intuitions would be in hypothetical circumstances. For example, when someone is presented with the option of killing one person to save many, as is the case of the famous trolley thought experiment, it does not teach us anything about the world, but it allows us to explore the implicit consequences of our preexisting beliefs.

In Galileo's thought experiment refuting the Aristotelian scenario because of its inconsistency, the so-called *reductio* stage applies because it shows that heavier masses cannot fall more rapidly than light masses, and a parallel thought experiment would show that light bodies cannot fall more quickly than heavy bodies. Thus, only the remaining option, that all masses fall at the same rate, can be the correct one.

Thought experiments may fail by being unable to answer the “what if” questions correctly or due to its Inconsistency. A thought experiment is more likely to succeed if explicit or implicit knowledge of the behavior of real phenomena can predict how hypothetical events would unfold. If a consistent model is achieved, one can conclude that the scenario envisioned is possible (Cooper, 2005).

1.5 _ Conceiving the Thought Experiment

In order to conceive a thought experiment, one has to know to what end it is created, and this implies having sufficient knowledge concerning its objective(s) e.g. a thesis to

refute or to confirm, but also about the kind of scenario envisaged. All this information has to be accessible in the thought experiment's narrative so that it can be played out accordingly with the specific details in its setup. The narrative entails the objectives and the specific characteristics of the scenario and purposely neglects various factors that can influence both the outside world as well as the thought experiment itself. So, in some sense it also gives us information about what we are not supposed to imagine. The narrative's tools are the concepts that guide us through the unfolding of the thought experiment hence the importance of the background knowledge to which they are linked. As Gooding argues, "A thought experiment becomes possible [only] when a world is sufficiently well-represented that experimental procedures and their likely consequences can be described within it" (Gooding 1992, p 281).

The thought experiment's narrative induces us to imagine the setup and we tend to unfold the course of possible outcomes with an anticipatory baggage of rules linked to the background knowledge which will then condition and restrict the imagined scenarios. To perform a thought experiment is to "live through" the anticipatory horizons against the background of which fulfilled intentions towards particular phases of the imaginative process always stand out. (Wiltsche, 2016). These anticipation rules are of course also influenced by the target-thesis that in some sense pre-selects what we should be paying attention to, and to what we should not be paying attention to. Wiltsche argues that thought experiments should be understood as anticipation pumps and that these help to outline how knowledge is generated in the realm of thought experiment reasoning. In this train of thought of his it would be fair to say that this makes the thought experiment's narrative all the more relevant since it triggers the right kind of background knowledge that in turn motivates the right kinds of anticipatory rules in order to reach the desired outcome. "Unless the imaginative context dictates otherwise, we tend to fill in the details of our fictional worlds with background beliefs about the actual world" (Walton, 1990, p144–61).

"One of the reasons why immersion is epistemically significant is that it gives us the opportunity to test our background knowledge under conditions that could not be replicated otherwise." (Wiltsche, 2016, p 21). Of course, there is also a danger in projecting quasi-worlds that are far removed from the actual world: The more outlandish the quasi-world, the higher the risk that our anticipations eventually become ungrounded. But one of the things a well-designed thought experiment accomplishes by way of

projecting idealized quasi-worlds is to provoke collisions between parts of our background knowledge that would not even come close under normal circumstances. By scrutinizing our background assumptions, thought experiments can explore that existing background knowledge and get rid of inconsistencies previously left unquestioned, and in some impressive cases even reconfigure the conceptual apparatus with which we usually approach the world. (Kuhn 1977; Gendler 1998).

Before presenting the thought experiments (3a, 3b & 3c) that I put forward below, and after having presented the main rules for devising a thought experiment and mentioning some of the most significant thought experiments of the past, an important question springs forth: How sure can one be that one's thought experiment is original? It is widely accepted that physical experiments can be replicated (although there can be differences), but to what extent is a thought experiment replicable?

An approach adopted by Brown (1991) is that as long as a continuity in the central narrative is maintained then thought experiments can be replicated by different people at different times. Historical research shows that just as it happens in empirical experiments, thought experiments too can be replicated and even remastered for different purposes at any time as was the case of David Bohm's modification of the EPR Gedankenexperiment I mentioned earlier. Naturally this prompts the question about the identity a thought experiment has when there are two similar narratives. Moreover, will it be possible to reach different conclusions out of the same thought experiment?

Roy A. Sorensen addresses this identity issue and, in his intuitionist approach the thought experiment is defined by a set of propositions that if changed (even if it is only one proposition that changes) then a new thought experiment arises (Sorensen, 1992). Also, Norton has this view that a modified thought experiment stands for a new one. In Nersessian's simulative model-based reasoning there is also room for various manipulations of these models where the interaction of various constraints can reveal some in detriment of others in a new, adapted construction of the model thus enabling them to be replicated or remastered (Nersessian, 2008).

For two thought experiments to be considered as the same thought experiment they must (1) have the same material realization; or (2) have the same theoretical interpretation; or (3) both the same material realization and theoretical interpretation. Prompted by some prominent historical examples, one might conclude that either

condition (1) or condition (2) is sufficient for them to count as the same thought experiment. (Bokulich and Frappier, 2018).

The three thought experiments designed for the purpose of this thesis aim to discuss specific subjects and issues such as identity, intelligence and consciousness. These are then in the subsequent chapters of mind uploading (MU - Part III) and superintelligence (SI - Part IV) duly explored where:

- (1) The thought experiment iterating human predictor response (TE-3a) will ultimately argue that through mind uploading the personal identity issue is still very much of an open question;
- (2) The thought experiment iterating human intelligence (TE-3b) will argue that an intelligence augmentation may be possible via mind uploading and through the emergence of a superintelligence. It is again addressed and analyzed in the (SSI-CA) chapter, since it was presented to the interviewees as a sub-question;
- (3) The thought experiment, the delusion of reality (TE-3c/d) will argue about the perpetuation of consciousness via mind uploading. Its possible outcomes are discussed in a metaphysical context supporting mind uploading's purpose and need as a technology, hence also supporting this thesis' context and reason altogether.

In the following section, a state-of-the-art review concerning the (NBIC) nano-bio-info-cognitive sciences and technologies will be described in order to contextualize the thought experiments' fields of interaction and better help to understand the scope of their applicability. The thought experiments will then in the last section of this chapter be presented.

2 _ Innovative technologies as a scientific support for mind uploading.

By the third quarter of the twenty-first century, \$1,000 will buy a computer a billion times more intelligent than every human on Earth combined. AI will surpass

humans as the smartest and most capable life-forms on Earth. (Louis Del Monte, 2013)

2.1 _ Artificial (General) Intelligence

In 1956 McCarthy, Minsky, Newell and Simon, motivated by advances in neurology, information theory, cybernetics and computer hardware and software, founded the field of Artificial Intelligence at a conference held at Dartmouth College. Between the mid 60s and the 90s, AI was heavily funded, and expectations exceeded results. After a so-called AI winter in which development seemed to slow down, at the beginning of the twentieth century resurgence of AI began through new applications in data mining and medical diagnostics amongst others. Due to the emergence of expert systems focusing on addressing well-defined specific applications, computational power was getting closer to that of a human brain so today AI is a commonplace embedded in every modern technology. It got imperceptibly into our lives even though it did not even exist a generation ago. The knowledge doubling curve is accelerating exponentially and using Moore's Law we can extrapolate that in terms of raw processing power (petaflops), computer processing power will meet or exceed that of the human mind by the 2020s which still does not mean we will have a computer matching the human mind. Software plays a key role in both processing power (MIPS) and AI (Del Monte, 2013). The difference between AI and AGI, lies in thinking machines programmed to be problem solving and task orienting mechanisms (narrow AI) with general-purpose systems with an intelligence that can be compared to that of the human brain (AGI). (Goertzel, 2012).

2.2 _ Intelligence Explosion

Intelligence measures an agent's ability to achieve goals in a wide range of environments (Hutter, 2012). Some call this the "optimization power/used resources" concept of intelligence. AI refers to "systems which match or exceed human intelligence in virtually all domains of interest" (Shulman and Bostrom 2012). Humans may create human-level artificial intelligence this century and shortly after we may witness an intelligence explosion or technological singularity a chain of events by which human-level AI leads quite swiftly to intelligent systems with capabilities way beyond humankind (Muelhauser & Salamon, 2012). Once an AI system with roughly human-level general intelligence is created, an intelligence explosion involving the relatively rapid creation of increasingly

more intelligent systems will very likely follow, resulting in an emergence of dramatically superhuman intelligences (Goertzel, 2015).

2.3 _ Molecular Nanotechnology

NanoBiotechnology narrows the gap between technology and life with its interface. As this borderline tends to disappear so increase hopes (e.g. overcoming disease) and fears (e.g. loss of individuality) for human enhancement and thus reinforcing the debate over the concept of humanity (Grunwald, 2007). Life's essential building blocks such as DNA and proteins take place at the nanoscale and their manipulation renders tools able to transform cellular systems into controllable bio-nanomachines. The bottom-up approach (as opposed to top-down fabrication methodology) also goes through molecular nanotechnology and its natural download of atomic precise manufacturing (APM) providing force for a fourth revolution, and like the preceding ones, it will transform the structure of society (Drexler, 2013). Since we are just now starting to build things from zero, just as nature does (and has been doing for millions of years), we can expect even more precision tools than the ones we developed so far. Nanotechnologies promise to enable fundamental life processes since they thrive on the idea of manipulating atoms, hence biomolecules with precision. The nano-assembler, when or if fully operative, should be able to build anything even itself which would by itself signify a singularity in the sense that it would stand for a major disruption to humankind (Drexler, 1986). But aside from controversial hyped sci-fi-like breakthroughs, others quite feasible are also expected soon. The biomedical potential for enhancing humanity is huge but from healing and repairing to enhancing lies just a thin borderline and classical medicine could soon redesign the human body.

On the other hand, all these processes entail ethical issues that must also be addressed. Indeed, ethicists have warned about enhancement since the appearance of genetic engineering in the early 80s. Instead of treating diseases, gene therapy may be used to design human beings with a reduced need to sleep, increased life spans, or an increased ability to remember (Walters & Palmer, 1997). Today nano and biotechnology are applied in health care to prevent and cure diseases. However, transhumanism indicates the use of nanotechnology to transgress the limits of humankind by removing aging or enhancing intelligence (Ebbesen et al, 2006). Enhancement technologies stemming from genomics

have triggered the most controversies and deadlock between bioconservatives and transhumanists. Selection of embryos for or against certain genetic traits has arisen together with the threat of eugenics (Koch, 2010). But if genetic engineering was (is) about manipulating genes, nanotechnology is about manipulating atoms, hence a more profound impact is expected to happen and therefore more effective enhancements and dramatic modifications of the human genome.

Human sensory capabilities have been technically extended over the past years and will through nanoelectronic neuro-implants still be ever more in the near future. Damaged or deficient biological functions can be compensated through functional augmentation or replacement (cochlear and retinal implants). Interfacing the brain with these emergent technologies opened ways to establish interconnection to artificial limbs (prostheses) and expand conventional body functions. Reconstruction of technological equivalents of natural body functions gives way to enhancement as a natural step further. In the long run, one goal in human enhancement is to delay or even abolish death (Grunwald 2007).

Electronic circuits linked directly to the electrochemical activity of the brain also show progress of the interface between biological material with nanomaterial, thus making the abovementioned borderline evermore thinner (Preston, 2005). Nanotechnology based applications for repair and reconstruction of the central nervous system through nanoparticles specifically designed as vehicles to the CNS is also gaining momentum (Cho, 2012). Molecular nanotechnology will eventually enable us to selectively examine and repair cells, which will lead to considerable life extension. We will have the blueprint for cells and bring them back to a fully functioning youthful profile. Moreover, we will be able to change the original blueprint substituting carbon fibers that are harder than bone, or *respirocytes* that are upgrades to red blood cells for carrying oxygen. Nanotechnology will eventually reduce the cost and increase the resolution of brain scanning devices. We will be able to back up and reproduce our biological memories on faster and more robust hardware platforms (Jacobstein, 2017). Shaping the world atom by atom (Bensaude-Vincent, 2009) prompts expectations of reshaping the human body at the molecular level and of curing and repairing any illness. Knowing DNA to be the building block of life by excellence, then one can only imagine what can be done in the next few years toward enhancing the human and expanding over its biological boundaries.

2.4 _ Whole Brain Emulation (WBE) and Mind Uploading (MU)

Mind uploading is a popular term for the process by which the mind would be transferred from its original biological brain to an artificial computational substrate. This possibility gives rise to an eventual eternity in a digital form which is the reason for its great popularity among life extension supporters. An upload can be the result of a gradual replacement of brain parts or a scanning. It can be a destructive uploading with no preservation of the original brain, a nondestructive or even a reconstructive one from records of the cognitive structure. This raises a number of issues that will be assessed in detail in this thesis further ahead. Assuming uploading of the mind and everything it stands for is possible in technical ways (by WBE for instance), what would be transferred? Would it be an indistinguishable copy (numeric identity)? Is a copy the same self? Does it transfer a personal identity as the kind that preserves continuity of consciousness of the original self (continuation of subjective experience within the same entity through time)?

The majority of the suggested methods for creating or developing superintelligence are either via intelligence amplification of our natural human brains (IA), or through the progressive development of artificial intelligence (AI). Both of them represent a technological singularity in the sense that they entail a disruption of continuity in human evolutionary terms. There is a variety of ways speculated to achieve an intelligence augmentation, from bioengineering, neurotropic drugs and AI-assistants to brain computer interfaces (BCI) and mind uploading (MU). Furthermore, multiple paths to an intelligence explosion make a singularity more likely to happen; for a singularity not to occur they would all have to fail. The probability for such a singularity to occur therefore increments with the variety of technological solutions developing. Many authors claim that there is more potential in achieving this superintelligence with a non-biological substrate such as a computer than through biological amplification, and that humans even if enhanced, will ultimately be outclassed anyway. (Bostrom, 2012, Bostrom, 2014).

The emergence of superintelligence via the AI scenario prognosticates a future with digital minds and foresees the probability of AI turning into an existential problem for humankind (Sandberg, 2014) or *the worst thing to happen to humanity in history* (Hawking et al 2014). In this scenario, as will be discussed in more depth in sections ahead (namely Part IV, Superintelligence), a dystopian outcome might become very likely in our society, whereas in the other scenario of IA, merging technological

enhancements with human biology would result in an utopian outcome of humankind overcoming disease and aging. (Kurzweil, 2005). Either way, in both scenarios or possible outcomes from the human point of view, there would always be technical ground for mind uploading and this would by itself stand, for a variety of issues such as personal identity and the consciousness conundrum, as presented and discussed herein.

One of the main technological preconditions for mind-uploading techniques to take place depends highly on the feasibility of a Whole Brain Emulation (WBE). Exploring the Brain using computer technology to integrate everything and providing data about everything on a molecular level is a quite recent enterprise in our history so it should not shock us that the brain itself remains our greatest mystery and decoding it eventually is our grand achievement. The HBP (Human Brain Project) entails a number of research studies at different levels of organization and development and aims to upload the entire mouse brain this decade and parts of the human brain still in the coming years of the early 2020s. The ability to collect, store and analyze information, which allows us to transcend our biology, has increased a lot recently due to progress in computing. Exascale computing is arriving (2020) and is expected to perform at the cellular level in order to deliver simulations of the human brain in the near future. It is also expected to achieve a molecular-level simulation of parts of the brain when required. (Sandberg and Bostrom, 2008). This approach of working backwards from measurements of the functioning system to engineer models of how that system works is called reverse engineering.

Artificial neural networks are being developed to simulate how the human brain works. Reverse engineering will show how neurons connect, rendering us a virtual interaction, which will be confirmed by in vivo and in vitro experiments. Software simulating human brains at various levels could lead to WBE as long as entailing functional similarity regardless of exhibiting mental states or not. The current roadmap for the field suggests such emulations as viable by mid-century and thus boosting development of neuroprosthetic devices (Sandberg & Bostrom, 2008). Neuromorphic chips already in development will take the Human Brain Project to a level of replicating cognitive capabilities (Markram, 2013). Sandberg has thoroughly discussed the feasibility of a WBE as well as the criteria by which emulations are to be validated. In order to obtain a structural validity for a WBE to be considered possible, even in the long term, only simulations totally equivalent in terms of functionality would be accountable. This is the same as to claim that all relevant properties of the original system would have to be fully replicated (and observable, too, of course). (Sandberg, 2013).

2.5 _ The Consciousness Conundrum

One of the major reasons for our human natural suspicion towards an upload comes from the fact that we put great value in consciousness.

To some extent, our consciousness represents our main trace of personal identity, and without it, we wouldn't be ourselves anymore. However, we still cannot explain consciousness or how it can emerge from the brain. Dennett dedicated a number of years in researching consciousness and as a renamed cognitive scientist and philosopher, claims that consciousness does not "reside" in the brain nor does it "flow" at all. To him, the so-called stream of consciousness is just an illusion and there is no evidence that it exists all the time. (Dennett, 1991). In Häggström's thought experiment on the issue, humans base their judgement in other people being conscious or not on their speech and behavior rather than on their anatomy (Häggström 2016). In this manner we might be tricked into thinking an upload would be conscious by behaving as a conscious being, just like Alan Turing claimed more than 50 years ago when he considered machines could become intelligent and because of that we might be tricked believing they were so. It should then be reasonable at least to assume that any emulated system could have the same material properties as the original one so we should treat it accordingly (Sandberg 2014). The CToM (Computational Theory of Mind) states that it is not the material substance that matters for a conscious state to take place, but rather its organizational pattern, i.e., the right kind of processing, as will be thoroughly analyzed and discussed more ahead in the consciousness section of the Mind Uploading chapter of this thesis (Part III). (Häggström, 2016).

But this doesn't help us in the personal identity issue since personal identity is not an organizational invariant. A clone, or even an identical twin, would still be a different person, numerically, even though qualitatively they could even be identical at the molecular level, we still would have two different identities. This would mean that even though an uploading could preserve an organizational state such as consciousness this would still not mean it would preserve a personal identity. So, replacing every single molecule would not be enough. Even if fundamental physics were to allow it, which doesn't seem to be the case, particles do not have or transport identities and the atoms we are made of keep changing over time but we still remain the same, thus it aren't the individual molecules themselves but the organizational pattern which persists over time. The identity theory (mind and consciousness are identical to brain states) fell to the

arguments of functionalism which claim that function and causal structure of the brain are responsible for consciousness (Chalmers 1996, Levin 2013). There is some evidence pointing to the cognitive structure of the brain which we might call the information pattern that really counts instead of the physical matter itself, the substance, as will be addressed again in more detail in the coming sections (Part III, Mind Uploading) as mentioned earlier. Hence, a machine that would copy the cognitive architecture of the human brain might generate consciousness just as the original system does. (Cerullo, 2015).

2.6 _ Concluding Remarks

If consciousness is in fact an organizational invariant independently of the substrate, whether that organization is implemented in neurons or in some other substrate as silicon for instance, then it should be able to be replicated in a computer simulation and generate a conscious state as the human brain itself. (Chalmers, 2010).

Depending on the feasibility of whole brain emulation (WBE) or a neural network emulation, either by gradual replacement of brain parts or by scanning, mind uploading would be achievable and substrate independent minds (SIMs) could come real. As humankind progresses in time towards a post-human Era, technologies seem to fuse efforts in transforming our main biological entity into an eventual non-biological substrate containing our essence. Our identity would in a digital form be preserved. Or maybe not. This dilemma directly relates to the consciousness conundrum which in turn opens a new debate around humanism as exposed and discussed along this thesis.

3 _ MU + SI Thought Experiments

“If there is a way of guaranteeing that superior artificial intellects will never harm human beings then such intellects will be created. If there is no way to have such a guarantee, then they will probably be created nevertheless.” (Bostrom 2005)

[3a _ iterating human predictor response](#)

[\(this TE finds its application / discussion in the MU chapter regarding the identity issue\)](#)

CaseScenario 1 _ AI development leads to WBE, hence mind uploading.

CaseScenario 2 _ Avatar competes in different scenarios.

CaseScenario 3 _ Uploads iteration enables predictive response.

CaseScenario 1 _ SIMs comparative algorithm & the identity issue.

There is, for the time being, and will be even more in the near future, a strong and increasing pressure to improve AI up to human level.

1 _ Consider the exponential technological progress to sustain its pace in the forthcoming years up until whole brain emulation (WBE) becomes feasible at the human brain level. Uploading of one's brain capacities and characteristics would at this particular moment constitute a mere technicality since one thing leads consequently to the other. So, if up until then ethical guidelines and the political agenda allow it, one would copy the data & information processor we call the mind to generate a digital copy as an avatar of oneself. Would it copy one's identity? This is one of the major philosophical concerns and the one that fuels some skepticism regarding WBE. Let us put aside, for the time being, the question on whether self is in fact you, your best copy, your best conscious copy or even just a technical software version of what your mind stands for without demonstrated identity nor consciousness. Just an emulation of yourself. We'll get to that later.

2 _ Now consider your copy to be in a digital form, reproducible and also editable, just as happens today with any kind of digital file, be it photographic, video or audio, text or processing sheet. Remember that you can multiply these files as many times as you wish and in an immediate fashion too as opposed to several generations taking place in genetics and keep in mind that it will always be 100% equal to the original file you created. You would then be able to run this file, your avatar, in other compatible programs. Now consider that by then we have already designed an algorithm where your avatar can be measured and confronted in various different environments in order to find out the response your avatar(you) make(s) toward specific adversities, actually your response in case you would have undergone the experiment yourself. But you cannot be submitted physically to such an experience because it could kill you or because it would take place in a hypothetical environment that is not reproducible today either for being unethical or far too expensive or even physically impossible.

The evolution of humankind shows us that the prediction of events, whether naturally occurring or not, ingrained us with a leverage over other species. In competitive environments foreseeing upcoming events has over time proven to be crucial for humankind's survival. Consciousness may have developed in our brains precisely because of this. Forecasting empowers us with an adaptive advantage that enables us to tackle upcoming obstacles or better deal with cyclic forthcoming events and basically keeps us a step ahead of whatever it is that comes our way.

3 _ You would now have your best possible response to specific conditions without actually or physically going through them. You wouldn't have the time or the stamina to undergo such a procedure yourself as stated before. But your avatar would. Making an iteration of this avatar submitted to a million different hypothetical scenarios would provide us with an optimal and practical human response towards specific conditions: for instance, making it ultimately possible for us to avoid disasters or crises of any kind.

4 _ Now let's take a step further toward the identity issue. Imagine you would now compare the blueprint from your copy with someone else's copy. For the sake of argument, let us call your copy SIM#01 (SIM being substrate independent mind), and the copy of someone randomly else chosen as SIM#99. Now SIM#99 would be, just as its genetic sequence would, in principle somewhat different from your SIM#01. You would definitely expect your SIM#01 to respond to the set of scenarios exposed somewhat differently than SIM#99 would. But what if, after a millionfold conditions submitted (and the number of conditions would only be limited by the computational power in terms of database and memory capacity), the difference between SIM#01 and SIM#99 would be narrowed to 1%? Could we infer that, even at this stage still without knowing whether or not your personality, consciousness and identity have actually been transferred, you and SIM#99 would definitely share 99% of a cognitive blueprint, hence 99% of all that it stands for, e.g., personality, consciousness and identity?

Now imagine the mindblowing case that in fact, SIM#01 and SIM#99 would not only share that same cognitive blueprint, but also respond this time around to the different scenarios exposed above with an astonishing 100% identity. Would you (unfortunately for you and for SIM#99 too) not be able to claim that you were actually no different from SIM#99 and that you and SIM#99 share the same identity, given that although both of you are different persons, you would actually prove to be mathematically, statistically and scientifically identical?

These scenarios are actually in reference to humankind although individually addressed to as copies, i.e., the copy is an example of a whole group inasmuch as being representative of the human species like a sample. Hence, one should regard the experiment's outcome as a potential benefit (if such were to be obtained) to humankind as opposed to a personal one.

3b _ iterating human intelligence

(this TE finds its application / discussion in the SI chapter regarding the intelligence issue)

CaseScenario 1 _ AI leads to superintelligence.

CaseScenario 2 _ Multiple uploads compete to select the most intelligent copy.

CaseScenario 3 _ Uploads iteration avoids singleton.

1 _ Consider artificial intelligence has developed a lot but not quite yet as much as to be considered a superintelligence of some kind, so it still hasn't reached that threshold we today mean by technological singularity.

2 _ Now consider changing the variables of the thought experiment 3a. Swap them and instead of iterating the conditions, we would maintain the same scenario your copy would compete in. Only this time you would submit millions of copies of yourself running against the same problem confronted. But given that we can edit your copy as we like, we'd make small changes starting from copy#1 @ copy#1million. We could then run a plot and select the best performing copy.

3 _ We are driven to seek for a propagation in a cosmic scale because of natural selection, so would we not be mainly preoccupied with the survival of our own kin and offspring? If not for any other reason, is it not fair for us to claim that protecting and assuring the survival of our offspring leads to endurance and propagation of ourselves even in individual terms? Aren't our offspring the closest we get in guaranteeing survival of at least part of ourselves? Then uploading, even in a destructive way, even without preservation of identity (our offspring do not preserve our identity either) would surely account for at least the second best way to propagate us, if not the first one. On the other

hand, *en route* to that propagation, and assuming we keep on developing AI until it develops itself into something eventually superior to us, wouldn't we by then turn into direct competitors of AI? But we can only remain so if we can match the pace of its development, only if we can keep up with it. Beyond that we'd be naturally selected to lose the driver's seat in evolution.

The competitive effects of selection are ever-present in what may be called Universal Darwinism (as coined by D. Dennett in 2005), so toward a hypothetical technological singularity either we would have already merged with AI in order to avoid being sidestepped, or we would develop the technology to upload us, if not with personal identity than at least copies capable of doing everything we won't do for any reason and capable of doing things we cannot imagine yet today. You might then even download your upgraded self/avatar (SIM) reversely back into your brain. This iteration of intelligence in order to excel, to have the best performance could keep us updated towards an eventual AGI intelligence explosion and avoid a bad technological singularity outcome such as the development of a Singleton of some kind.

As a side note, It is important to notice that for ethical reasons other authors believe it to be more admissible to experiment on copies resulting of only specific parts of the brain, responsible for specific functions, therefore, in principle, not conscious parts, or parts not being associated for consciousness, rather than trying out iterations with uploads resulting from whole brain emulation, where consciousness would or could be emulated. This does not make much sense to me, since there is no such way as to measure consciousness, whether in a whole brain or not, it for sure would not mean that a region or part of the brain responsible for vision would not entail any form of consciousness, simply by the fact of being specific. Actually, this could be quite debatable, since it could be responsible and specific only for vision, it doesn't mean for seeing; as the part responsible for hearing, wouldn't necessarily be responsible for listening.

3c the delusion of reality

[\(this TE finds its application / discussion in the MU chapter regarding the consciousness issue\)](#)

CaseScenario 1 _ Emulated reality subdues real reality.

CaseScenario 2 _ Humankind dies out / post-human consciousness takes over.

CaseScenario 3 _ Endless simulations & the tradeoff.

1 _ Mind uploading would enable an augmented intelligence result of an AGI enhanced human brain. This superintelligence would be able to improve virtual reality into a level outclassing reality itself. Sensations would be hyped, experiences excelled and life exponentiated. Along with enhanced intelligence would come another level of consciousness and post-humans would realize that the former reality would actually only be a simulation we've been led to think of as being real. After mind uploading became common sense, and after technical issues got completely surpassed inasmuch as the virtual realm played out in our uploaded minds became so real, humans would start to change into uploads altogether.

2 _ Humankind would die out as a biological homo sapiens because nobody would want to live that life anymore. Post-humans would only exist in the digital realm from that moment on with this new level of consciousness. However, ultimately, we would even determine life to have always been played out this way in a repetitive kind of iteration of various realities and actually none of them being true, doomed to be imprisoned for eternity in endless levels of simulations. But the extra consciousness also gave us an edge of awareness of knowing how to escape this plot of simulations cycle to *the other side*. This *other side* might be *the real reality*, but one that we couldn't grasp as human biological finite beings.

3 _ Consider that there was a tradeoff and a glitch in the system, a kind of black hole, an opaque singularity that could be an escape out of it. The tradeoff would be that in order to avoid the doomed scenario we would have to give everything up to restart, from the big bang so that evolution would unfold again from the very first particle onwards. Without knowing if it would evolve in such a way as to develop human beings but knowing the next time around there would be a possibility of evolving into some kind of entity capable of getting out of this simulation hell. We would then have the choice, the tools and the means to reverse the whole thing. We would have developed the necessary technology to simply go back to *ordinary life* as it was experienced before we were aware of this, back to non-virtual reality and start over.

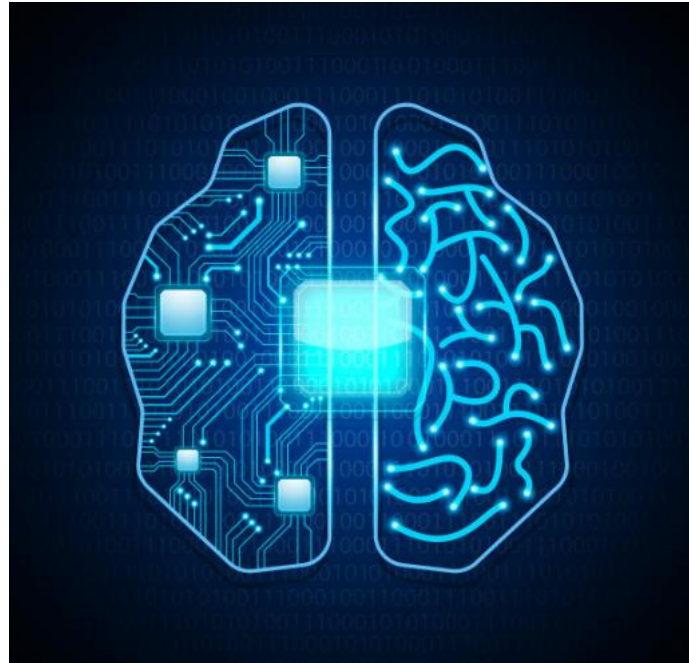
Now, if we decide to escape the system, human beings as *human beings* would have a low probability of developing. All the conditions and the course of events would have to be exactly the same to increase those chances. The trade-off would most certainly mean

human extinction. On the other hand, it could mean that a new kind of entity resulting from that new evolution could be able to figure out the glitch, escape the simulations and enable life to become real? Maybe taste *the meaning of life*?

The real question is if we get a consciousness level that labels our reality as untrue but lends us the awareness of a true reality yet in a non-human condition; do we choose to cease human existence to embrace the chance of experiencing *that real existence*? Or do we choose the seemingly hedonistic paradise of living forever although knowing it to be a *false reality*?

What do we value the most? Being human? Or *just* being?

PART II _ Whole Brain Emulation



*§ markers indicate the linked paradigm in context that can be found throughout the chapter.

“Whole brain emulation (WBE) is the possible future one-to-one modeling of the function of the entire (human) brain. The basic idea is to take a particular brain, scan its structure in detail, and construct a software model of it that is so faithful to the original that, when run on appropriate hardware, it will behave in essentially the same way as the original brain. This would achieve software-based intelligence by copying biological intelligence without necessarily understanding it.”

in Feasibility of whole brain emulation in Vincent C. Müller (ed.), Theory and Philosophy of Artificial Intelligence (SAPERE; Berlin: Springer), Sandberg, Anders, 2013

1 _ Intro

In the nineties, neuroscientist Randal Koene joined a mailing list called the mind uploading research group (MURG) formed by Joe Strout who was known for having practical ideas about the reconstruction of brains by building compartmental models of neurons from structural scans. It was during that period (April 29th, 1998) that Koene coined the term Whole Brain Emulation (WBE) for reasons of clarity, and as he claims, “to move from good ideas to feasible projects” (Koene, 2015).

Science fiction of course had a big influence in those early days, which actually also led to some skepticism among various neuroscientists together with the fact that it was mostly discussed in a philosophical realm until it started to gain traction in scientific surroundings. There have been many thought experiments among various renown philosophical authors such as Searle (1980), Parfit (1984) and Chalmers (1995) exploring brain emulation as a tool for arguing mind and identity theories without taking stock of the technological feasibility issue. The philosophical debate, although absolutely essential and desirable, at this early beginning of WBE as a discipline deviated it from its real scientific intentions and held scientists and their research a bit off track. It makes more sense to address the philosophical issues regarding consciousness at a time when the underlying high-level functions involving emulations would already be sufficiently understood and modeled to better satisfy the philosophical criteria for qualia and consciousness matters.

By the time the first WBE workshop convened at Oxford University in 2008, and as a result of a technical report on its feasibility, the first roadmap regarding WBE was drawn by Anders Sandberg together with Nick Bostrom (Future of Humanity Institute, Oxford University). At the Singularity Summit of 2009, Koene and others were able to take WBE into consideration as a transformative technology that should be contrasted with Artificial General Intelligence (AGI) and its potential risks. At the annual conference on AGI in Lugano (Switzerland) in 2010, WBE was included for the first time, and again, Koene and others revamped the organizational network around the multidisciplinary research and development efforts toward WBE. Since then several projects with the goal of large-scale scanning and reconstruction started off, such as The Human Brain Project, The Blue Brain, The Human Connectome Project and The Brain Preservation Foundation. These will be addressed below.

In that roadmap of 2008, Sandberg and Bostrom summarized the participants' views on several relevant issues toward the WBE goal keeping technical requirements and constraints in mind. They began to outline a number of reasons for which it should be considered an important academic matter, as follows:

- Research

- o Brain emulation is the logical endpoint of computational neuroscience's attempts to accurately model neurons and brain systems.

- o Brain emulation could help us to understand the brain, both in the lead-up to successful emulation and afterwards by providing an ideal test bed for neuroscientific experimentation and study.

- o Neuromorphic engineering based on partial results would be useful in a number of applications such as pattern recognition, AI and brain-computer interfaces.

- o As a long-term research goal it might be a strong vision to stimulate computational neuroscience.

- o As a case for future studies it represents a case where a radical future possibility can be examined in the light of current knowledge.

- Economics

- o The economic impact of copiable brains could be immense, and could have profound societal consequences (Hanson, 1994, 2008b). Even low probability events of such magnitude merit investigation.

- Individually

- o If emulation of particular brains is possible and affordable, and if concerns about individual identity can be met, such emulation would enable back-up copies and "digital immortality".

- Philosophy

- o Brain emulation would itself be a test for many ideas in the philosophy of mind and philosophy of identity or provide a novel context for thinking about such ideas.

o It may represent a radical new form of human enhancement.

Just like Koene underlined differences to AGI in the Singularity summit mentioned earlier, Sandberg and Bostrom also stressed the need for it at this roadmap. They considered WBE to be an exceptional engineering and research problem and with a well-defined goal that could be reached by extrapolations of current technologies whereas in what concerns AI, we still do not know how far we are from success since we don't have any clear metric of it (even today in 2020).

More recently, in March 20, 2014, Randal Koene gave an interview to Luke Muehlhauser pinpointing some differences and updates in relation to the roadmap established in 2008 following the aforementioned conference on WBE. As it turns out all of the involved at that venue pretty much agreed that that roadmap did not actually lay out concrete / well devised theoretical plans toward WBE's feasibility. He conveyed the idea that the roadmap from 2008 focused on structural data acquisition instead of giving relevance to the system identification issue that actually is the fundamental problem in an unknown black box kind of system (as is the brain). Thus, according to Koene and this roadmap update, this all hinges on making a plan of decomposing WBE into feasible individual parts.

As it seems, in depth analysis of the brain shows that at different levels we find smaller black boxes, and this stands for a big challenge in neuroscience: that of combining large-scale and high-resolution data to better understand functional changes in the brain. These kinds of issues regarding available technologies and the level of precision and accuracy one wants to achieve led neuroscientists dedicated to WBE to precise the kind of data and models that would actually be relevant for WBE. Hence the importance in system identification.

Bearing this in mind a new roadmap of the technological requirements to achieve WBE through system identification was postulated in 2014 in order to break the whole problem into smaller subsystems that communicate with each other instead of approaching the whole brain as a unique unknown system that could fail in capturing much latent function. System identification is somewhat disclosing the process by which an unknown system (like a black box, like the brain itself) turns input into output (Ljung, 2008). It is possible to identify a complete series of input patterns by observing a working system. So with smaller subsystems it is possible to better manage them individually, in identifying the

problems they expose in a single manner, enabling an identification that also better suits the available specific tools to measure and collect data, as well as to better build functional models to ultimately devise prosthetic replacements. (Koene & Deca). Summarizing the strategy, we can say it would start by choosing the subsystems (1) to then disclose their connection (2) and the way they communicate with each other in order to finally enable identification (3) of their systems' functions. According to Koene et al, it is by addressing this three-step subsystems' identification strategy separately, that a roadmap toward a feasible WBE is doable: "Experimental Research in whole brain emulation aims to re-implement functions of a mind in another computational substrate. This will be achieved by carefully emulating the function of fundamental components and by copying the connectivity between those components. The necessary precision must enable prediction of the natural development of active states. In order to accomplish this prediction in-vivo measurements at large scale and high resolution are critically important. We propose a set of requirements for these empirical measurements. We then outline general methods leading to acquisition of a structural and functional connectome, and to the characterization of responses at large scale and high resolution." (Koene, 2012 *β*1, p. 35-65). Finally, some project developments that tackle the problem of functional recording in-vivo, as the "molecular ticker-tape" and others are presented in the final sections of this chapter.

The ultimate objective of a WBE, insofar as it involves a precise functional reconstruction of neural tissue, is to achieve a reimplementation of the human brain similar enough to behave like the basic computational functions carried out by elements of the neurophysiology as well as the connectivity existing between them, in the neuroanatomy of the original brain. It entails a sequence of events regarding detailed information about the data contained in the original brain and the technologies necessary to scan and analyze its structure. Among the main requirements are the following challenges: 1) The amount of neurobiological information necessary; 2) The extent to which the reimplementation would have to be similar; and 3) The criteria used in order to accurately evaluate the success rate of that emulation. In terms of technology these requirements can somewhat correspond: 1) Scanning the brain; 2) Translating the scan into a model and 3) Running that model on a computer. These requirements will be individually analyzed below, after we clarify Whole Brain Emulation's goals, as its scope has been somewhat updated after revising strategies, available technologies and consequent roadmaps in these last few years.

2 _ WBE's Objectives

In the last few years it has become evident that even if its main objectives would fall short, WBE would have at least produced a significant amount of data available for the development of AI itself. Furthermore, models designed to do simulations *en route* to that WBE, would definitely provide insight at all levels of neurobiology, leading to a better understanding of the behavior of neurons to that of networks of neurons and ultimately to how the brain works overall.

As expected, many algorithmic approaches in developing AI are bioinspired. (§1) Hence, before AI reaches human level intelligence (AGI) it can of course progress a lot with WBE, since ideas from simulations of neural networks may help develop intelligent behavior in computers for instance through massive parallelism as already happens with neural networks used in various computer applications. Moreover, attempts to achieve WBE would carry byproducts such as to better develop and enhance prosthetics. (Cattell et al, 2012). “For the most part, computers still use the same basic architecture envisioned by John von Neumann in 1945. Hardware architectures based on the massive parallelism and adaptability of the brain may yield new computer architectures and micro-architectures that can be applied to problems currently intractable with conventional computing and networking architectures.” (Cattell, 2012, p2).

It would also contribute immensely to test ideas and thought experiments regarding philosophy of mind and of identity, consequently leading to that ongoing debate over the so-called hard problem of consciousness (and others). As observed in the 2008 roadmap, it also obviously stands for the logical endpoint of computational neuroscience to accurately model neurons and brain systems since it would deliver maps of biological systems at different levels of resolution. WBE would be a combination of an accurate map and sufficiently accurate modeling. (Sandberg, *Feasibility of Whole Brain Emulation*, 2013).

On the other hand, WBE could lead to AGI and to superintelligence as I will argue below.

A successful WBE would have a tremendous impact in each and every field. As Hanson argues, “given that human capital is a main driver of economic growth, copiable human capital (in the sense of systems able to perform the same tasks as a human) implies

extremely fast economic growth and would have profound societal and ethical consequences” (Hanson, 2008, p45-50).

Sandberg claims that skepticism about WBE is not sufficient to refute it because we can find many past examples in which other complex systems such as genomes, proteins and integrated circuits were scanned or simulated successfully despite being judged as apparently impossible just a few years earlier. [\(§2\)](#)

As a concluding remark one could now trace two goals of WBE. The first being the acquisition of data representing the neuronal network of a specific human brain and the second one being the recreation of that particular neuronal network by a computer emulation. The resulting byproducts as well as brain preservation and preparation efforts, as neuron and glial cell behavior and neural network understanding and modeling, all have a direct bearing on WBE. (Feierbach & Koene, Whole Brain Emulation - State of the Art, 2015).

As a result of these last developments (ca. 10 years, up until 2020), huge funding efforts have been dedicated both in the USA and the EU (EFPL 2014, NIH 2014) that will be duly addressed further on in this thread in the analogical linkage section dedicated to some of the main ongoing projects. [\(§4\)](#)

3 _ Technical Requirements

3.1 _ Scanning the brain (and the amount of neurobiological information necessary)

One can define the amount of neurobiological information necessary at the level of understanding, the level of detail we would have to have about the brain in order to perform a successful scan, which is tied to the so-called scalability issue. At present there (still) is no consensus on what level would be needed. “An informal poll among researchers suggested that the electrophysiological level (cellular compartments) is most popular, but this merely represents an opinion” (Sandberg & Bostrom, 2008, p9). There is a need for scale separation between smaller and larger functional units of the brain because if it does not exist then WBE might well be infeasible. No method with a

finite size cut-off would achieve an emulation, which makes this one of the biggest uncertainties of brain scanning for WBE. Scale separation would enable to model the larger parts of the system with their rules, which is to say with their own constraints that need to respect the laws of physics without worrying with the smallest parts of the system. Interactions on smaller scales must of course be part of the overall state of the emulation in order to validate that particular simulation insofar as their macroscale outcome has to be captured as well. (Sandberg, 2013).

Nevertheless, there are encouraging examples in science, including systems involving both quantum mechanics and classical physics where the behavior of those quantum particles when in large numbers in aggregate, can be described by classical physics. (§3) So the classical interactions can be modeled without knowledge of the quantum interactions. One would expect a similar scale separation to exist in the brain since one can also develop rules for neurons and/or ion channels to operate without knowing their interactions at the atomic level. (Daniel Eth et al, 2014). There is evidence of physical systems that show scale separation. There exists a level where interactions on shorter length and time scales average out, producing macroscale dynamics uncoupled from the dynamics on smaller scales. (Hillerbrand, 2008). Sandberg gives the statistical mechanics of gases as an example (as he also did in his interview for this thesis when addressing this issue), that in order to describe the behavior of the system at the macroscale, it is not absolutely necessary to know its exact molecular interactions. (Sandberg, 2013; Sandberg, 2020). “While the scale separation between different levels of the nervous system does not have a radical separation as in statistical physics, the different levels of the hierarchy – neural fields, neuron populations, individual neurons, ion channels – are often separated by one or two orders of magnitude, and may hence be amenable to statistical treatments that average small and fast scales, possibly introducing random noise from the coarse graining of micro-dynamics (Berglund, 2011).” (Sandberg, 2013, p260).

Actually, at the synaptic level (or higher), it is likely to occur a scale separation since the environment in which the brain operates is known to be noisy, with external (e.g. environmental conditions, microtraumas, changing nutrient states, parasites etc.) and internal disturbances (e.g. developmental noise, thermal noise, chemical noise), happening on a regular basis and that do not seem to affect the brain’s normal functioning. These are therefore most probably interactions that do not interfere with the

overall performance of the brain otherwise if they were to be fundamental interactions, then noise would have an impact on the overall system's performance. (Daniel Eth et al, 2014). A scale separation is likely to exist at higher levels than those of specific concentrations of neurotransmitters and of metabolites, which is where the amplitude of this noise is. Furthermore, since some neurons can communicate even without spiking, it seems probable for a scale separation to not exist above or at this level of spiking neurons. (Lytton, 2002). Hence, error correction and robustness mechanisms are expected to exist at this level of 'non-interference'. As Sandberg recalls, "brains sensitive to microscale properties for their functioning would exhibit erratic and non-adaptive behavior. If the differences introduced by simulation are smaller than the normal noise level (and of correctable type) then it is likely that scale separation would occur." (Sandberg, 2013, p261).

Following Koene's view, a range of different levels of detail will be involved in WBE. From a functional perspective it might be possible that an emulation at the level of the spike timing and that of patterns of neural spikes is the level perceptible for us. "Consider, differences of pre and postsynaptic spike times are the basis for synaptic strengthening (spike-timing dependent potentiation), i.e. encoding of long-term memory. Trains of spikes are used to communicate sensory input (visual, auditory, etc). Patterns of spikes are used to drive groups of muscles (locomotion, speech, etc...)" (Koene, 2014, interview given to Luke Muelhauser for the Machine Intelligence Research Institute newsletter, 132). According to Koene, a good emulation would probably require a deeper level of data acquisition for parameter estimation which makes him believe that WBE will ultimately involve a combination of tools that could retrieve data at one level but that could then also go at deeper levels in those places or at those times, that would require a more local dynamics. Moreover, as other authors (e.g. Eth, Sandberg, Hayworth) he does not think it would be necessary for us to map all molecules throughout an entire brain. Although, to his reasoning, it would be likely for us to actually study their dynamics inasmuch as to understand how to relate measurable quantities with the parameters to be represented in the emulation. (Koene, 2014).

Small voltage deviations in digital integrated circuits are still treated as zeros and ones because it does not affect the resulting outcome inasmuch as small deviations in membrane potentials of neurons have no effect on the action potential generated in the postsynaptic potentials. There are of course many more examples like these that

basically show that by analyzing the capacity for error correction we can identify scale separation and that there is proof of biological systems that resist the discretization noise through these error correction mechanisms that prompt discrete attractor dynamics, which is in accordance with synaptic plasticity also involving discrete changes instead of graded responses. (Ajay & Bhalla, 2006; Elliott, 2011). Hence we can infer about a possible scale separation at the neuronal and synaptic level as Sandberg claims: “Information is transmitted in a discrete code (with a possible exception of timing) between discrete entities. At finer resolution thermal and chemical noise will be significant, suggesting that evolution would have promoted error correction and hence scale separation.” (Sandberg, 2013, p262).

A successful WBE would definitely prove that all relevant processes have in fact been included in the simulation but on the other hand, if this would not be the case, then analyzing failed attempts, especially in conjunction with in vivo studies, would provide information about missed modalities, thus also being a valuable amount of information and data, as previously stated.

3.2 _ Translating the scan into a model (and the extent to which the reimplementation would have to be similar)

The brain most certainly has chaotic dynamics (Li, Yu, & Liao, 2001), meaning that the state of the simulation would also most certainly diverge from the original brain, which also would mean a low kind of predictive validity of that same simulation. But brains exhibit the same kind of behavior, inasmuch as they still are recognizable as being the same, even when at slightly different activity states so it would appear that causal structure as opposed to the exact dynamic state, is what actually matters for a successful reimplementation. A WBE to be structurally valid, does not require an output between simulation and modeled brain to be identical, but it does demand a compatible output with the one that the brain would have given if it were in a similar state. (Sandberg, 2013).

After scanning the brain, the following step would of course be that of enabling it to be run on a computer. There would be a kind of template, a broad framework, a model of the human brain to work as an input to generate the parameters, in such a way as to

allow the data from that particular brain-scan to fit in. This template, or model, would not have to contain prior knowledge or information of each level of function of the brain, but would have to be based on the functionally relevant behavior of the components used for computing and connectivity between those components. (Koene, 2012).

As stated above (in the scanning the brain section above), a scale separation is likely to happen at a higher level than that of the neurotransmitter and metabolite concentrations level, more like at the synaptic level or even higher. Yet, it is also true that genetic defects that eliminate specific neurotransmitters, such as dopamine for instance, affect dramatically a person's behavior which of course aligns with their concentration's importance in maintaining the brain's normal functionality. It would then make sense to model that above mentioned template with a finer resolution than just that of the synaptic level since these substances, including ions and even chemicals, are the ones trading value conditioning the synaptic activity. (Daniel Eth et al, 2014).

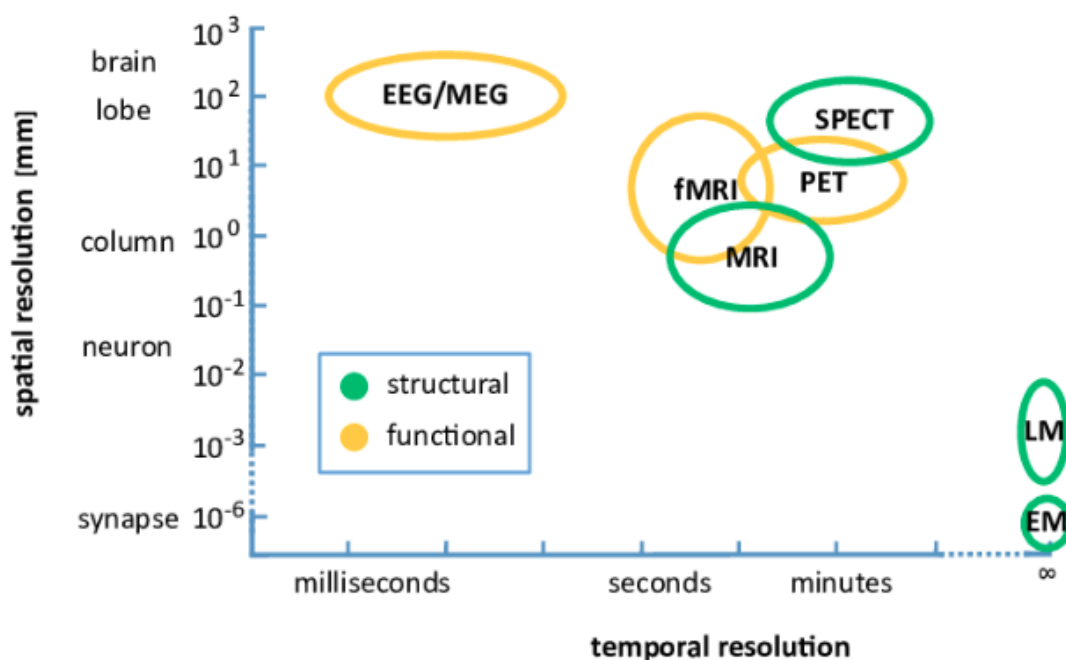


Figure 02 – Spatial resolution of various brain scanning technologies versus temporal resolution, showing the differences between functional and structural applicabilities of them. (Source: Pfister, et al, Visualization in Connectomics, Mathematics and Visualization · June 2012).

As one can easily see in Figure 02, and as the pervasiveness clearly proves, the more spatial resolution the traditional available technologies have, all the more invasive they

are, ultimately meaning that they become destructive. Destroying the brain in the process of scanning it does not allow us to obtain in vivo real-time data of the functionality, although being highly adequate to obtain structural data of the brain. Brain scanning technologies are traditionally divided between being noninvasive such as the cases of functional Magnetic Resonance Imaging (fMRI), magnetoencephalography (MEG) and positron emission tomography (PET) [yellow in the plot], that are adequate for measuring the brain's activity (functional) and invasive (the destructive ones) such as the case of EM (structural) [green in the plot].

Eth argues that “activity might be obtainable from structure once we have a general model of the brain, but in order to create this general model, we will need to determine activity to discern how the molecules interact in vivo.” (Eth, 2014, p134). Also, from the plot one can determine that both the poor temporal resolution of fMRI and the poor spatial resolution of PET can be compensated by adding electroencephalography (EEG) and MEG, although these are lacking in resolution when compared with electron microscopes. Although these brain scanning technologies have developed quite a lot in the last years, even exponentially in the case of MEG and MRI as can be seen on Figure 03, it does not seem that the desired spatial resolution can be achieved in the coming few years with noninvasive nondestructive techniques.

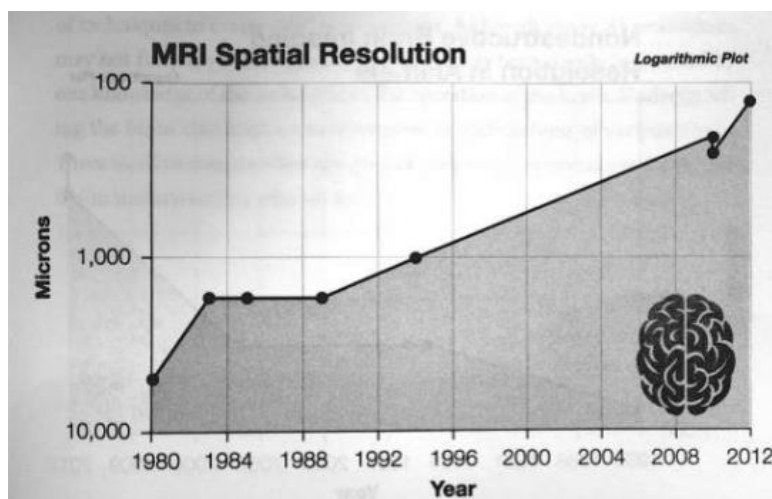


Figure 03 – Spatial resolution of Magnetic Resonance Imaging (MRI) for use in brain scanning, in microns. Since the y-axis is logarithmic and the curve is more or less linear, the resolution has improved more or less exponentially. (Kurzweil, 2012) Unfortunately, even if this trend continues for the decades to come MRI still will not have the spatial resolution needed to see pieces of neuroanatomy relevant to WBE (including synapses). (Source: Kurzweil, 2012).

In order to build that template (a general model) the spatial resolution has to improve insofar as to provide data at the synaptic level which still is a couple of orders of magnitude short. On the other hand, the invasive technologies of course seem to fulfill these spatial resolution constraints. There seems to be, even now, sufficient capacity in imaging volumes of neural tissues to distinguish the finest fibers to obtain an unparalleled detail (Hayworth, Kasthuri, Schalek, & Lichtman, 2006).

However, given that these invasive technologies are destructive this would of course constitute an ethical issue, so again it does not seem to be acceptable even though it has the desired kind of detail level. The research necessary for a brain emulation to succeed, will surely involve aversive conditions and scanning the brain will likely be a destructive process. (Sandberg, *Ethics for Brain Emulation*, 2014). But, to some extent we can bypass some of those ethical issues in the sense that neuroscientific experiments have been carried out using brain models other than the human brain. A wide range of various different organisms' brain models have been employed, thus not involving human brains in the process, and ultimately also to some extent depending on the specificity of the techniques used, i.e., some models are more adequate than others varying in genetic, structural and functionality studies.

Genetic information is obviously always important. And for WBE it can be used by manipulating the brain's functional connections in order to understand specific neural systems as would be, for example, the case of the dopaminergic system. This would elicit valuable functional knowledge to enable an eventual WBE without necessarily understanding the entire structure. (Deca, *Available Tools For WBE*, 2011). The fly's brain is an adequate model for this, since drosophilas have been widely and successfully used in genomics. Hence, to design a better model without the human brain, taking advantage of genetic manipulation of the brain's plasticity regarding neural architecture can be done and is of the utmost importance toward a successful WBE.

On the other hand, the mouse brain model has already allowed for in vivo imaging at single cell resolution with electron and 2-photon confocal microscopy, which therefore permits in vivo imaging of neuronal networks, hence activity dependent functional changes occurring in these networks, can be observed and monitored. Furthermore, the mouse brain is of course much closer to the human brain, thus whenever possible, also depending on the technical requirements available, these are preferable models to the abovementioned ones. Once sliced, one can even correlate structure and function in

these brain models, by combining in vivo and in vitro imaging. These data showing the correlation between structural and functional changes in different types of cells may then be obtained from 3D models with in vivo imaging of the mouse brain, which again would have been impossible to carry out on human brains with the current technologies.

These data may contribute in accelerating the mapping of the mouse brain, with connectivity in response to different stimuli. It will then, further on, be correlated with other models to explore functional and structural differences and to ultimately apply in WBE. In terms of circuitry, optical stimulation and techniques such as TMS help analyze the brain as if it were one big circuit and therefore connecting structure to function without having to comprehend all other processes, which as duly noted previously, may be important in achieving WBE. (Deca, 2011).

Other techniques have also been developed at the neuronal level such as the patch clamp, that permits studying in vivo electrical properties whether in neurons as well as in ion channels. These techniques have meanwhile enabled models of individual neurons, although not large circuits of them, which will be the ultimate goal: To create neuronal circuits by the time the aforementioned template, or general model, is to be made and eventually containing millions of neurons. (Alivisatos et al, 2012).

It is possible to emulate specific functions and understand their evolution through the combination of various techniques with different types of acquisition and analysis tools, which stands as an advantage for WBE. Also, for different physical levels (which range from the physiological to the biochemical level of analysis), there are different conceptual levels to which different approaches are applied. There also is, already, an immense volume of information regarding ion-channels within neurons and of their electrophysiology affecting their connectivity, which then also affect their plasticity ultimately resulting in differential brain functions. (Deca, 2011).

But what seems to lack is a goal-driven interdisciplinarity. Deca claims that “in this case the goal of WBE includes the goal of connecting structure and function, as well as high resolution with large scale simulations.” (Deca, 2011). As different studies have been carried through by different teams, also with different goals, there seems to be urgency in interfacing different stakeholders assembling knowledge and expertise of different kinds together toward a common objective, which could exponentiate the goal of a WBE.

[\(§4\)](#)

As Bostrom and Sandberg stressed out in that first roadmap mentioned above, we do not need to know everything to make a WBE possible but bearing in mind the goal of a WBE, I might add that - we do need to know what we need to know. And then connect and assemble these different inputs and insights toward that one specific goal.

3.2.1 _ The Connectome

The open connectome, as the name indicates, is an open source platform database available for anyone to access freely and to contribute with neural data ranging from cellular (micro) to fMRI gathered data at the macro level, as well as the types of analysis that can be performed with it. Its goal to completely reverse-engineer the human brain, containing up to date an impressive* record of anatomical (mouse) and functional (fMRI) data, resembles other mapping projects humankind has made in the past, as the Human Genome Project. (§2) One of the bottlenecks for the connectome and other large data projects is the enormous need for fast retrieval of data and storage capacity. Through the likes of automation of neuroscience on one hand, and data mining and machine learning algorithms on the other, retrieval of data might speed up a lot in the coming years.

Automation of neuroscience has been making progress and does apply a lot to traditional brain scanning methods. MIT researchers developed a method for retrieving data from neurons in living brains, which was previously performed by an expert. The authors claim that automation of this method is actually faster and more accurate than humans did before. (Trafton, 2012). Also, quite recently, biomedical engineers developed an automated process with AI and deep-learning algorithms that map and trace the shape of active neurons as accurately as humans but of course way much faster, thus condensing hours or even days of work in just a few seconds. The AI allows researchers to rapidly gather and process neuronal signals on real-time studies. (Somayyeh Soltanian-Zadeh, 2019). To generate a model of the human brain, these powerful data processing approaches have, nevertheless, to become even more powerful, at least for the purpose of a WBE to be feasible. These algorithms are of course also expected to improve with more powerful hardware capabilities, which in turn are, as mentioned earlier, also improving and at a high rate (if not exponential). This issue will be addressed again more ahead in the Analytical Inferences section to better justify the plot therein discussed.

Mapping the whole brain from neurons through synapses, to functional connectivity between different areas, aims to turn the connectome into a milestone toward WBE. At some stage, to carry out such a project has to entail a very big level of automation to keep up with the enormous amount of data, which will also probably involve data mining and machine learning, which on the other hand will need huge funding, making it a critical enterprise. (§4)

“The Human Genome Project generated \$800 billion for a \$3.8 billion investment. (Tripp and Grueber, 2011). If these projects follow a similar path, future investments in mapping the brain could be even bigger. Countries might realize the economic benefits of these projects, from drug discoveries to AI improvements to WBE. If an “arms race” of brain mapping takes off, funding for neuroscience could skyrocket.” (Daniel Eth, et al, 2014).

*<https://neuroscienceblueprint.nih.gov/human-connectome/connectome-programs>

3.3 _ Running that model on a computer (and the criteria used in order to accurately evaluate the success rate of that emulation)

If all properties and internal causal links of the original system have been replicated, then an equivalence of complete functionality would have been achieved with the simulation. But in order to validate this, one has to know exactly what it is that was supposed to be replicated in the first place and therefore, in the case of the human brain, what kind of requisites there are that fulfill this. These requisites are contingent to the criteria we define for a successful goal, and to the resolution scale required in the brain.

Validation depends upon testing if the real and the simulated systems correspond to each other. There are three types of validity (Zeigler, Praehofer, & Kim, 2000):

- Replicative validity: the simulation matches already observed data from the real system (retrodiction).
- Predictive validity: the simulation matches data before they are acquired from the real system (prediction).

- Structural validity: the simulation not only reproduces but reflects exactly how the original system operates. Every connection and process of the real system has total correspondence in the simulated one.

WBE itself, by definition, has to deliver a validity of the structural kind. But since we don't even know exactly how the brain works in every aspect, and as already discussed, WBE could still be achieved without having to know everything about the brain – then only a replicative kind of validity could be verifiable.

In computing, emulations occur when the functions of a system (whether hardware or software) have been duplicated from one to another. “Typically, the focus is on exact reproduction of external behavior rather than the exact internal structure. Internal states need to change as a function of inputs, producing outputs compatible with the modeled system but the states are not necessarily corresponding to any components of the system.” (Sandberg, 2013, p253). Still, in computing, successful emulations happen on a regular basis whenever a new computer runs the software of an older one and is able to emulate the underlying operating system of the older hardware and simulate it on a one-to-one manner. The emulation is an interface that is sufficiently accurate to imitate the so-called low-level functions which are not relevant for the simulation to take place.

Likewise, in the human brain we can recognize a similar kind of computability. When we make a digital remastering of an audio recording, some information gets lost. It is, nevertheless compensated inasmuch as one cannot distinguish the actual difference, since the general behavior of that sound is pretty much the same, though not infinitely precise. This digitalization is therefore a very good emulation of the analog original sound record and comparatively we could assume that WBE would not require replication at an infinite precision. There seems to be evidence of a kind of intelligible cooperation between regions of the brain as Koene argues: “In essence, the brain uses tricks like spike bursts, nested oscillatory modulation, redundant ensembles of neurons active in spike patterns, etc, to ensure reliability and be more computable.” (Koene, 2014, 132).

The key factors in emulation are the functional relations occurring in the original system, in the sense that they have to be transferred even if not entirely understood (even in the original one). But even in this case, not fully understanding them, at least they have to be acknowledged, recognized in their role, inasmuch as they have to unfold again in the emulated system, as they did in the original. This is why WBE is invariably tied to

functionalism. The relevant properties that functional relationships carry out in the original brain have to emerge from their execution in the emulated copy as well, otherwise it does not make much sense to talk about a WBE anyway, at least not in its holistic definition. However, as I claimed earlier in this chapter, WBE as a project seems to be justified anyway, because its efforts tend to lead to many valuable results whether scientifically, philosophically or even as an engineering goal.

Anders Sandberg filters out some potential success criteria based on the kind of emulation that could be observable. As already discussed, a successful emulation would always have to produce a similar emergent behavior as the original source system. If it would be sufficiently similar to that particular one brain, it would be an individual form of brain emulation, providing that this pattern would exhibit an exclusive behavior different from other brains. If not exclusive, then it could be a more generic form of emulation. If it wouldn't exhibit a specific behavior at all, but nevertheless an emulation that would produce a substantial range of similar emergent activity, then it would be a form of functional brain emulation. If a continuation of the original mind would be the case, either as numerically the same person, or as a surviving continuer thereof (Sandberg, 2013), then it would definitely count as a personal identity form of emulation. However only the first three could be directly observable in the sense that they do not depend on metaphysical assumptions. In the latter one, the criteria would never be absolute in the sense that it would depend upon the kind of theory of mind and identity subjected. (Chalmers, 2010).

If we take into account other criteria for validating the success of a WBE as goal-oriented and tied to system identification, then we could redefine WBE as "Brain emulation strives to achieve a functional re-implementation by which it is possible to predict an active brain state and behavior at a time $t + dt$ (with acceptable error) if we know the state at a slightly earlier time t ." (Koene & Deca, 2014, p2).

3.3.1 _ Non-organicism

Given that the feasibility of a WBE does not depend upon completely understanding how the neural system works, only to copy its internal functioning, this meaning we do not have to fully know the brain, rather only recognize and copy its functional interactions, then – as some authors have been underlying – whole brain emulation's approach to AI poses a key challenge: How to really validate the chosen criteria against a system whose

function is unknown? This has been called non-organicism (Sandberg, 2013, Koene, 2014). In order for WBE to take place, similar to a so-called strong AI (maybe even already AGI) that merely needs to exhibit an intelligent behavior (in the wider sense as systems that act like they have minds), a mind emulation might not be required: only the appropriate behavior. (Sandberg, 2013). Actually, can we measure our level of understanding? Or is this the measure of understanding? We would only need to understand the low-level actions that build the system, not the system itself.

Going even deeper in this challenge, we might even compare a WBE with an AGI for all matters. This is exactly what Koene holds: “If we accept that definitions of generality and of intelligence used in AGI can apply to human minds, then a reproduction of the processes of a human mind via whole brain emulation is a type of AGI” (Koene, 2012b p6). Until quite recently (ca. 20 years ago) very little was known regarding the underlying mechanistic architecture and functionality of the brain and scalability has been a challenge to neuroscience ever since. The development of various technologies (such as fMRI, PET and MEG) somewhat diminished this problem but it still remains a big hurdle as already exposed in the previous section. Nevertheless, great advancements have been announced and the rate of them is growing, along with that of exponential technological progress. [\(§2\)](#)

In the introduction to this chapter I mentioned the system identification approach by Randal Koene in updating the (08) roadmap of a WBE at that time. It seems appropriate to address that approach again now, and also a third time by the end of the chapter again, for an update of that roadmap is happening right now (2020). This reflects the progress WBE has made, in particular in this last decade, almost as technology itself, a field of specific interests that might in the near future even become an independent field of study on its own. Following Koene’s roadmap of 2015 toward WBE, there is supposed to be a momentum, a kind of threshold to reach when those hurdles mentioned above are surpassed. Once we develop the tools for high-resolution measurements to solve specific individually tractable system identification problems, we can turn emulation of the human brain’s scale into a feasible project.

1 _ This then involves four steps starting off with an iterative process of evaluation of proofs-of-concept so to speak, which Koene calls Hypothesis testing. These proofs-of-concept are a sort of experiential criteria according to well-defined performance requirements and to the experimental goals previously established. (Koene, 2012a). An

example of such an experimental performance being successful is the bio-mimetic chip of Berger's team entailing an experimental hippocampal neural prosthetic. This is a proof-of-concept in the sense that, according to Koene "a transfer function was identified and used to replicate the operational properties of biological neural circuitry in a region of the rat hippocampus known as CA3. In experiments, the prosthesis is able to reproduce the way in which input to the region is turned into output from that region." (Koene & Deca, *Journal of Artificial General Intelligence*, 2014, p1). The proposed systematic approach to brain emulation embodied by a neuroprosthesis produced successful experimental results in cognitive prosthetic work. (Berger et al, 2011).

2 _ The 2nd step would be to gather data (for instance also the data out of the connectome), and divide the system identification into smaller problems, as mentioned earlier in the previous section, into subsystems. This decomposition of the system step is relative to structure. The interest in connectome data has led to rapid tool development already, and great results have been reported in the labs of Winfried Denk (Max Planck), Jeff Lichtman (Harvard) and Ken Hayworth (Janelia Farms). (Koene et al, 2014). These relate to results obtained with volume microscopy where images are the product of ultra-thin layers of brain tissue in which individual synapses have been identified at a resolution of 5-10nm allowing reconstructing individual neurons' 3D geometry with detailed morphology of axon and dendrite branches. (Koene et al, 2014). This kind of level of detail would satisfy authors who think that identification of small axons, dendrites and synapses is the level needed to acquire relevant structural knowledge. Although eventually not needing to know everything about the brain in order to emulate it, we will definitely need to know the basic building blocks (which means to identify and understand neurotransmitters, synaptic junctions, neurons and their connectivities, the influence of glial cells, etc.) and how to functionally simulate them.

3_ After this structural scanning (connectomics) comes the step relative to function, which would be defined by the characterization of the system, where development of specific tools for the functional recording takes place, which actually can be done by addressing specific initiatives such as the BRAIN initiative (Obama, 2013). There have been promising tool developments based on reverse engineering the brain-based approaches such as looking for means of taking measurements from the inside, from within. Again, biomimicry plays an important role as an approach to retrieve data: In order to extract the necessary functional data from sub-systems, detection can take place in

physical proximity to the sources of interaction, i.e., at the core and through microscopic synaptic receptor channels. This approach would also fulfil spatial and temporal resolution requirements; hence better characterize functionality in vivo. (Koene et al, 2014). DNA is also being used as a kind of recording media of functional events which can then be retrieved from the cells.

4_ The fourth and last step would of course be that of emulation, which Koene et al define as the mathematical representations and computational platforms needed. WBE would become infeasible if there are un-computable functions in the human brain's regular activity (at least on conventional architectures). However, there has not been, up to date, any convincing empirical evidence for it and on the other hand there has been proof of biological computability. (Sandberg, 2013). Nevertheless, if there are un-computable operations happening in the brain's components then even if WBE could tackle this problem by maybe building unconventional hardware architectures (which could involve quantum computing [§3](#)), this would still constitute a fundamental issue: How would we keep track of the simulation to validate an occurrence of WBE? I will argue below that the computing power to overcome any kind of computing-related issue can become available (the question being more when than if). [§2](#) + [§3](#)

“Using the iterative approach described here, based on rigorous system identification and decomposition into feasibly characterized sub-systems, a neuroprosthetic reproduction of a mind may be created via whole brain emulation in the coming decades.” (Koene & Deca, Journal of Artificial General Intelligence, 2014, p5).

4 _ Analytical Inferences

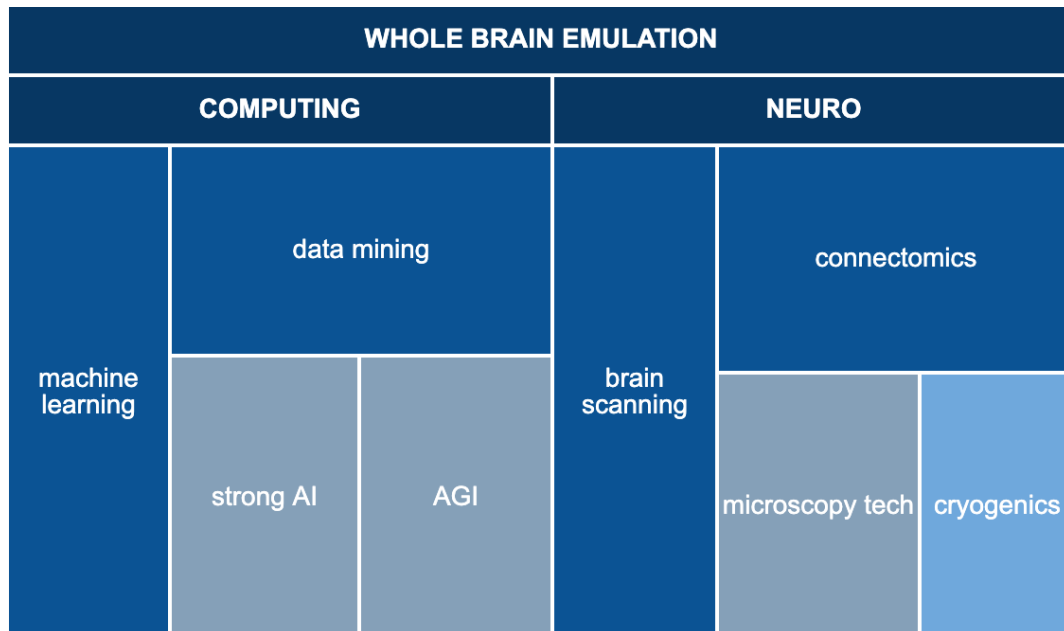


Chart 01 – WBE and its main fields of interaction analyzed in the plot A and plot B (Figure 05 & Figure 06), as they devise essentially in two big areas. The computing field with its own actors and the neuroscientific field, also with its own main topics of influence on WBE. (Source: Luig, 2020).

As it was shown in the previous sections, there are various critical issues toward a WBE that condition its feasibility. To achieve WBE every requirement that has been here described and analyzed has to be met. We can estimate a general trend toward that goal by evaluating each critical relative probabilistic strength measuring the factor of its requisite's relevance by its reliability (*).

- (*) Relevance X vs Reliability Y = WBE Feasibility

The values attributed in the evaluation of each critical issue according to the chosen criteria have taken into account three (3) levels of subjectivity tied to my personal knowledge on the subject. This is why the most subjective criteria have a lower confidence level, hence a bigger error bar associated in the plot generated. The thorough calculi made for this graph are available in a datasheet that can be found in the appendix of this thesis ([Appendix O](#)). The following is a reprise of its proceedings in order to advance to the analysis of its outcome.

The critical issues analyzed were the following:

- 1) Connectomics _ (NBIC-factor / multidisciplinary)
- 2) Brain Scanning _ (scanning techniques / microscopy resolution / scalability)
- 3) Computing Power _ (ongoing growth / high autonomy)
- 4) Non-Organicism _ (validation criteria / what to compare to?)
- 5) AGI _ (exponential growth of computing / deep machine learning)

Values ranging between -10 and 10 (adimensional), were given to each critical issue based on the three levels of subjectivity in relation to each of the 5 criteria defined in this chart (below). -10 would in case of the reliability mean absolute uncertainty as to the variable's future development (as being completely unpredictable) whereas +10 would mean absolute confidence in its outcome. The same criteria was applied to relevance, in this case -10 respectively meaning that the variable (critical issue) would not have any effect whatsoever in WBE's feasibility and conversely +10 being absolutely necessary for a WBE to take place.

The empirical criteria assigned to each of the variables (critical issues) in the WBE feasibility plot were the following:

- Funding _ (politics and economy)
- Autonomy _ (independence of the critical issue in developing)
- NBIC _ (multidisciplinary influence of nano-bio-info-cognitive technologies)
- Hype _ (stakeholders / social & cultural interest)
- Experts _ (views taken out of the semi-structured interviews / specific plot on SSI-Analysis chapter, Figures 29 & 30)

4.1 Connectomics @ PLOT

One could call this issue the result of the multidisciplinary endeavor involved in the roadmap of a WBE. The connectome as a project targets a large amount of data and information and enables to link different stakeholders' interests and efforts together. Mixing the know-how of different fields would elicit a kind of NBIC factor to synergistically boost knowledge and consequently the development of new applied techniques. It can be particularly important to solve specific barriers in issues such as the scalability and

the development of specific tools to apply in the identification of systems. Neuroscientific knowledge has of course to be generated toward the feasibility of WBE. Its funding has been considerably augmented in recent years and it will be much more: if not for other reasons, then because of the incrementing prevalence of neurologic diseases and disorders. This trend does not seem like it will stop anytime soon which will also indirectly boost funding toward a WBE though with a higher focus on basic neuroscientific research than in large-scale brain mapping projects. More directly related to WBE, the EU spent \$1.6B over ten years simulating the human brain through the Human Brain Project starting in 2013. (Dvorsky, 2013) The US also started a human brain project (the BRAIN Initiative), which has been suggested to cost around \$3B. (Markoff, 2013).

critical variables: (NBIC + / hype + / experts + / funding +)

@ PLOT: Relevance (8) Vs Reliability (8)

4.2 Brain Scanning (Techniques & Technologies) @ PLOT

It is clear that brain scanning techniques are (still) quite far from a desired momentum in the sense that they cannot gather the required neuroscientific knowledge to produce a general model of the brain. They are not autonomous in the sense that their dependence on the development of other technologies is essentially leading to uncertainty although the medical and disease related neurological market stands as one of its strong driving forces. The scalability factor also weighs hugely in uncertainty since it is not clear, and may well not be in the near future, to what extent specific neurological knowledge is in fact necessary. Nanotechnology may elicit a major breakthrough depending on the level of scale separation occurring in the brain and the one needed to record data in vivo at a neuronal level (examples can be found in the appendix in a reprise of some of the latest groundbreaking nano-approaches). Neuroscientists Anders Sandberg and Randal Koene also expressed their worries regarding the scalability issue here in this thesis' semi-structured interviews. Although defining himself as a functionalist, Sandberg shows some pessimism when approaching this particular problem, considering it a key challenge to overcome toward an eventual WBE.

critical variables: (autonomy - / NBIC + / experts -)

@ PLOT: Relevance (7) Vs Reliability (-2)

4.3 Computing Power @ PLOT

This could eventually be the Joker of the roadmap. Whether in processing capacity or in memory capacity it is difficult to imagine this issue to condition negatively that trend. Even if growth of computing would slow down, it does not seem possible to halt to such a level as not to allow the necessary power for a WBE to take place. Even though we do not know what amount of computing power this will require, there is strong confidence that it will happen, to the point of making this more a question of when. Furthermore, the exponential processing power following Moore's Law is likely to continue in the years to come as will be further addressed in sections ahead. This exceptional growth actually started long before 1965 when Gordon Moore, a computer scientist co-founder of INTEL observed the relation between processing capacity and size decrease of its components, establishing the so-called 2nd law, also known to be the fifth paradigm in computing (processing power increases every 18 months and decreases the size in the inverse order). It is an empirical observation, hence not a physical law, which does not make it less true and it is still applicable today.

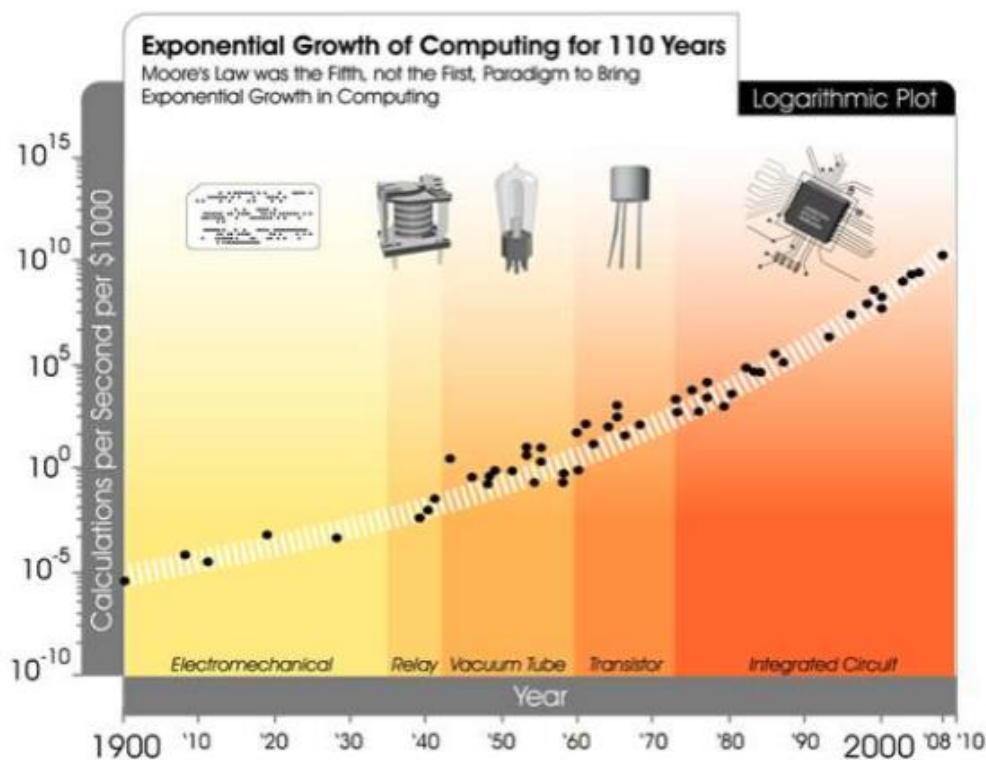


Figure 04 – Calculations per second (measured in floating-point operations per second) per constant \$1000 of hardware. Note that the y-axis is logarithmic, so the increasing slope indicates more than

exponential growth. This spans back over a century and will continue until hardware requirements for WBE can be available on computers costing no more than what PCs cost today. (Source: Kurzweil, 2010).

Calculations per second (CPS) measured in FLOPS (Floating Operations / Second) per constant dollars as indicated in Kurzweil's plots (Figure 04) have had an exponential growth over a century extending through various periods (paradigms) until integrated circuits and are believed to continue allowing WBE to be computationally feasible when the time arrives, meaning computers should have the power to simulate all the functional interactions happening at neurotransmitter level (even if not required) given estimates about the frequency in neurotransmitter transmission per synapse. (Sandberg & Bostrom, 2008).

Some uncertainty might arrive regarding the vast amount of data needed to generate a human brain model. But if computing processing power does not seem to represent a hurdle in time then memory capacity even less, since requirements for WBE at any scale one would want to emulate are so close to now as to claim that its growth would not have to be exponential anymore to reach the desired threshold. Moreover, Data mining & machine learning algorithms developed so far (and successfully in use in bioinformatics and stock trading) also seem to be on track to manage the amount of knowledge necessary. Furthermore, there is also a strong feeling in the computing science community that with improving hardware capabilities, also the algorithms will concomitantly improve. [\(§2\)](#)

There are several ways to estimate the gap between current capabilities and those required for a WBE. One of them is to compare computing trends to the number of entities/state variables needed to specify the emulation and the time resolution needed, as shown on the Table made for the brain emulation a roadmap of 2008. Sandberg and Bostrom, as mentioned right at the beginning of this chapter, made an extensive and thorough compilation of all relevant data regarding brain emulation. The computing power needed to achieve a WBE will depend on the level of detail. Thus, a molecular simulation would require an enormous amount of power (in a more distant future) whereas the neuronal level could be feasible within the coming years, of course always assuming trends like Moore's Law (and others) to maintain their pace. We can then estimate what kind of emulation is possible and how much it might cost. It is not possible to calculate with precision, but it is postulated that the human brain operates around

1exaFLOP. The human brain contains about 100 billion neurons and about a hundred trillion synapses. Each neuron can fire about 100 times a second. If we consider the brain as a neural network, then it would be equivalent to a machine requiring 10^{18} CPS and 10^{13} bits of memory. From June 2020, the Japanese Fugaku is the world's most powerful supercomputer, reaching 540 petaFLOPS and is meant to target 1 exaFLOP when finished (2021). The six-year budget for the system and related technology development is estimated around \$1 billion. Now, if we take the numbers estimated in that first roadmap for brain emulation (Sandberg and Bostrom, 2008), it was expected to reach 1 EFLOP for about \$1 million around the year 2019 (see Table 03 and Table 04, Sandberg and Bostrom, 2008). A rough comparison with today's Fugaku shows that the trend does not deviate that much and given conservative estimates in reach, supercomputer technology is expected to hit that scale in the next couple of years, assuming that the current pace of growth remains as it is. Thus, based on certain assumptions, and taking that roadmap's table as a reference, we can estimate when supercomputers might achieve the capabilities to emulate the human brain, about where a scale separation occurs. If what is needed for WBE is at the level of spiking neurons and synaptic states without actual ion concentrations (which is an optimistic stance), then 8 PB (8×10^{15} bytes) and 1 EFLOP/S (10^{18} FLOPS) are needed, which should have been available to a \$1M supercomputer by 2019 for both memory and processing power, as can be deduced from the Tables 02 & 03 (below). On the other hand if the desired scale should be that of the neurotransmitter concentrations (not as optimistic as the former one), then 1 EB (10^{18} bytes) and 10 YFLOPS (10^{25} FLOPS) are needed, which should be available to a \$1M supercomputer around 2029 for memory and 2044 for processing power. (Sandberg and Bostrom, 2008; Eth et al, 2013). Estimates are debatable, however, given that an order of magnitude complexity increase only adds circa 5 years to the estimate, adjustments should not affect the overall conclusion regarding WBE feasibility: "It appears feasible within the foreseeable future to store the full connectivity or even multistate compartment models of all neurons in the brain within the working memory of a large computing system." (Sandberg and Bostrom, 2008, p81).

critical variables: (autonomy + / NBIC + / experts + / funding + / hype +)

@ PLOT: Relevance (9) Vs Reliability (9)

Table 03 – Storage demands (emulation only, human brain). (Source: brain emulation roadmap, Sandberg and Bostrom, 2008).

Level		# entities	Bytes per entity	Memory demands (Tb)	Earliest year, \$1 million
1	Computational module	100-1,000?	?	?	?
2	Brain region connectivity	10^5 regions, 10^7 connections	3? (2-byte connectivity, 1 byte weight)	$3 \cdot 10^{-5}$	Present
3	Analog network population model	10^8 populations, 10^{13} connections.	5 (3-byte connectivity, 1 byte weight, 1 byte extra state variable)	50	Present
4	Spiking neural network	10^{11} neurons, 10^{15} connections.	8 (4-byte connectivity, 4 state variables)	8,000	2019
5	Electrophysiology	10^{15} compartments x 10 state variables = 10^{16} .	1 byte per state variable	10,000	2019
6	Metabolome	10^{16} compartments x 10^2 metabolites = 10^{18} .	1 byte per state variable	10^6	2029
7	Proteome	10^{16} compartments x 10^3 proteins and metabolites = 10^{19} .	1 byte per state variable	10^7	2034
8	States of protein complexes	10^{16} compartments x 10^3 proteins x 10 states = 10^{20}	1 byte per state variable	10^8	2038
9	Distribution of	10^{16} compartments x 10^3	1 byte per state	10^9	2043
	complexes	proteins and metabolites x 100 states/locations.	variable		
	Full 3D EM map (Fiala, 2002)	$50 \times 2.5 \times 2.5$ nm	1 byte per voxel, compressed.	10^9	2043
10	Stochastic behaviour of single molecules	10^{25} molecules	31 (2 bytes molecule type, 14 bytes position, 14 bytes velocity, 1 byte state)	$3.1 \cdot 10^{14}$	2069
11	Quantum	Either $\approx 10^{26}$ atoms, or smaller number of quantum-state carrying molecules.	Qbits	?	?

Table 04 – Processing demands (emulation only, human brain). (Source: brain emulation roadmap, Sandberg and Bostrom, 2008).

Level		# entities	FLOPS per entity	Time-steps per second	CPU demand (FLOPS)	Earliest year, \$1 million (commodity computer estimate)	Earliest year, \$1 million (Super-computer estimate)
1	Computational module	100-1,000?	?	?	?	?	?
2	Brain region connectivity	10^5 regions, 10^7 connections	?	?	?	?	?
3	Analog network population model	10^8 populations, 10^{13} connections	1	10^2	10^{15}	2023	2008 ¹⁵
4	Spiking neural network	10^{11} neurons, 10^{15} connections	10	10^3	10^{18}	2042	2019
5	Electrophysiology	10^{15} compartments x 10 state variables = 10^{16} .	10^3 FLOPS per ms per compartment.	10^4	10^{22}	2068	2033
6	Metabolome	10^{16} compartments x 10^2 metabolites = 10^{18} .	10^3	10^4	10^{25}	2087	2044
7	Proteome	10^{16} compartments x 10^3 proteins and metabolites = 10^{19} .	10^3	10^4	10^{26}	2093	2048
8	States of protein complexes	10^{16} compartments x 10^3 proteins x 10 states = 10^{20}	10^3	10^4	10^{27}	2100	2052
9	Distribution of	10^{16}	10^3	10^6	10^{30}	2119	2063
	complexes	compartments x 10^3 proteins and metabolites x 100 states per location = 10^{21} .					
10	Stochastic behavior of single molecules	10^{25} molecules	10^3	10^{15}	10^{43}	2201	2111
11	Quantum	Either $\approx 10^{26}$ atoms, or smaller number of quantum-state carrying molecules.	Qbits	10^{13} - 10^{20}	?	?	?

4.4 Non-Organicism @ PLOT

Assuming that a WBE would in fact be feasible, and performed, as discussed in the above sections, a question would still remain: How would and could we measure its success? Even with all the well-defined criteria already envisioned today to successfully evaluate the emulation process, unless we by then would also know how to measure and explain issues such as the hard problem of consciousness, this would constitute a major uncertainty and critical issue. Not as big as to slow down the processes of achieving a WBE, so not as much as to block WBE from taking place but maybe as big as to not being able to validate it successfully. Of course, and once again - the argument of not needing to know in order to copy and execute, can be applied here in this issue. We might achieve WBE but not validate it accordingly. Since total understanding of the brain is not required, only component parts and their functional interactions, we might in fact achieve emulation but not be able to recognize it. Taking Turing's Test as an example, a system intelligent enough might even trick us into thinking it is conscious, but still - it could be a simulation. We might not even know we have achieved it, because there is no way to measure it now, and it might not be any valid way to measure it by then, hence we might reach it without formally acknowledging we actually did.

Or would we be able to do it and it would then enable us, through an intelligence augmentation, to clear things out? This challenge will be accordingly addressed in the following chapters of mind uploading and superintelligence. Either way, non-organicism might not stand in the way of a WBE, but it still represents a major downforce of its concept, hence representing a big uncertainty.

critical variables: (funding - / NBIC + / experts - / autonomy -)

@ PLOT: Relevance (2) Vs Reliability (-8)

4.5 AGI (Artificial General Intelligence) @ PLOT

AGI and WBE have a lot in common. Both of them are evaluated in performance and neither implies the full understanding of intelligence. There is even an overlap of interests and areas in the research of these two fields of neuroscience and artificial general intelligence that underline the need for an interdisciplinary perspective (Goertzel & Pennachin, 2007). AGI and WBE can help each other out, but their goals are quite

different. WBE (or a neuroprosthesis for that matter) is successful if it efficiently replicates the original system's functions, which should be perceived and observed whereas an AGI would be considered successful when a machine would be able to perform at a human kind of intelligence level, which also means that it would have been able to capture the general principles of a mind. Hence their separate investigations suited accordingly to their goals.

Given that brain emulations are like black boxes /opaque systems), AGI might turn out to be an alternative to WBE *en route* to an augmentation of intelligence as long as we can fully understand AGI itself, as Anders Sandberg has expressed in his interview (in this thesis). It might be an easier task than that of developing a WBE. Because of eventual (in)compatibility issues and combining the human brain with AGI might render a very powerful system indeed, but it might also be a mountain to climb. Computer scientist Richard Loosemore argues about the motivation system stressing that its kind could define a friendly AI (or not) in becoming a controlled AGI. As also neuroscientist Kenneth Hayworth (president and co-founder of the Brain Preservation Foundation) underlined in his statements, again in the semi-structured interviews in this thesis (see chapter V/ questions III, IV), the fact that we never dealt with extreme differences in intelligence in our history, which could happen in a case of a rampant evolving out-of-control kind of AGI. Actually 9 out of 11 interviewees viewed the singleton scenario as quite possible. In a non-controlled kind of AGI development, a recursive overnight form of intelligence explosion would almost inevitably turn it into a singleton, which underlines the danger in a possible randomness of values regarding AGI. Hence, the uncertainties are very high for a number of different reasons.

critical variables: (hype + / NBIC + / experts -)

@ PLOT: Relevance (5) Vs Reliability (-5)

4.6. @ PLOTS

Both plots (presented below) relate the same data exhibiting the 5 critical issues that I consider relevant toward the feasibility of a WBE, and the same validation criteria used, only differing in their presentation. The distance between those more relevant and reliable requisites to the lesser ones is more dramatically visible in Plot A than in Plot B (Figure 5 & Figure 6). But relative subjectivity (grey bar) tied to the validation criteria

appears more enhanced in Plot B showing that even if these were to alter in a reasonable error window (error bars), it would not change the overall evidence that there are some negative driving forces mainly linked to their uncertainty (negative reliability) in estimating feasibility. This does not mean that a WBE is therefore more or less quantitatively probable of course, it just means that there are uncertainties that have to be overcome *en route* to a WBE. It is an estimative approach that helps identify and highlight pros & cons in WBE feasibility.

Computing Power and Connectomics may seem to be reasonably safe bets towards that goal. It does not seem very likely for computing power to become the bottleneck for WBE and human brain emulations should be achievable before the mid-century. But Non-Organicism, or the difficulty in establishing an indisputable validation criterion is certainly not on the plus side and even AGI may not necessarily work in favor of WBE feasibility because of its uncertainty by default. Traditional Brain Scanning techniques will of course improve along with technological progress altogether and with funding in neuroscience generally speaking since they share the same market forces, including medical imaging. However, their dependency on several other fields and technologies to be able to improve in order to ultimately gather all the necessary data by themselves, does add some uncertainty too.

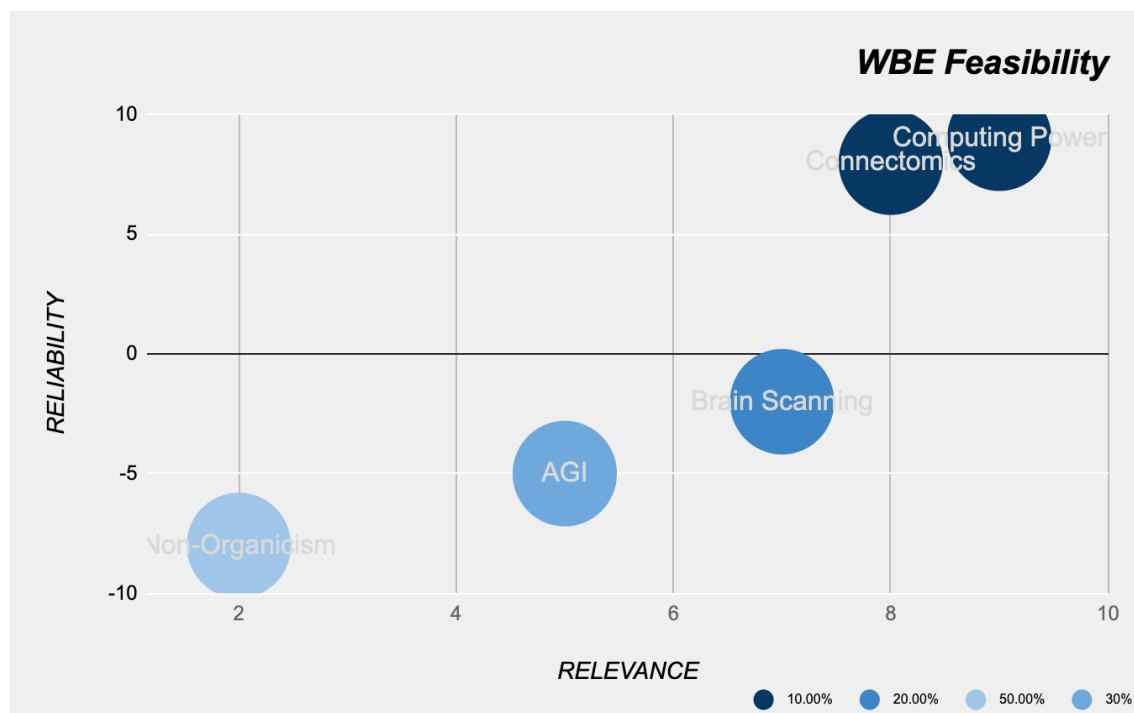


Figure 05 – (Plot A: WBE feasibility regarding critical issues. (Luig, 2020).

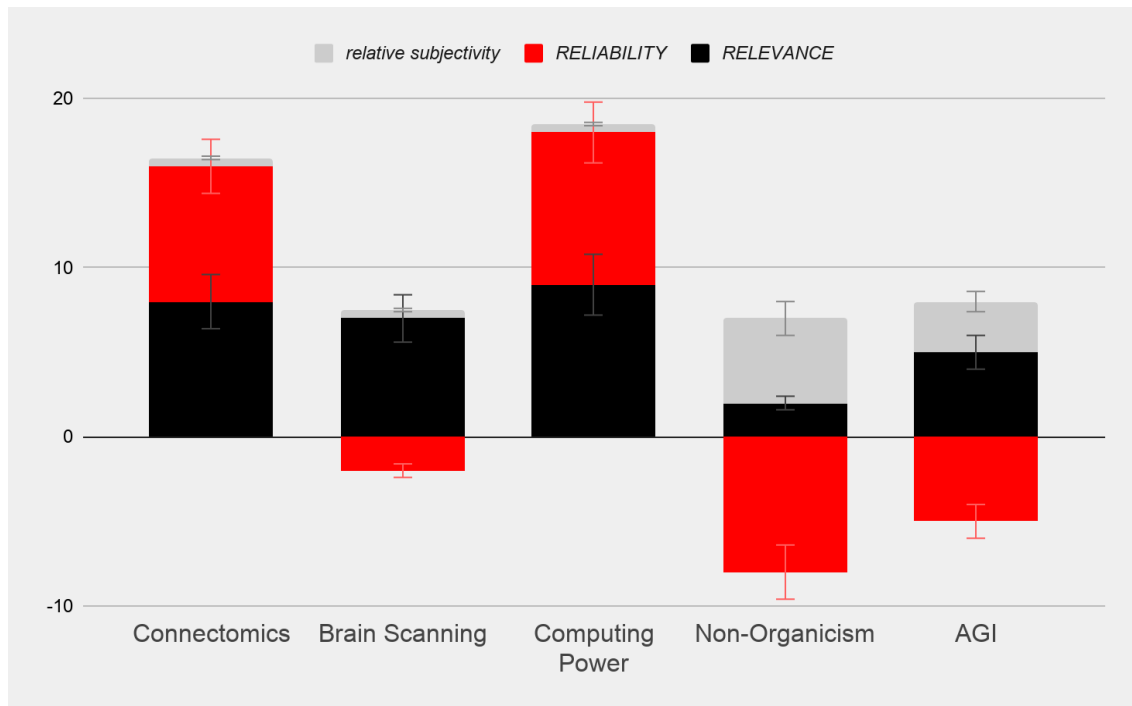


Figure 06 – (Plot B): WBE feasibility regarding critical issues. (Luig, 2020).

5 _ Preliminary Concluding Remarks A

If exponential gains in processing power would come to an end in the near future, then the necessary computing power for a WBE to take place will not suffice in order to emulate a brain in the absence of a scale separation in the brain. I will discuss this pessimistic outcome in the following chapter (analogical linkage).

In the subsequent section some arguments favoring an eventual WBE will be made using paradigmatic concepts in the sense that they do change events and shift both the mentality when addressing critical issues as well as the pace of progress. I call this an analogical linkage because they establish analogies with different fields and different approaches taking advantage of that association of thoughts. Some of the state-of-the-art examples used in this thread are of course applicable in different sections too. It does not change their argumentative power, also because the paradigms are to some extent transversely immersed everywhere and in everything.

Koene described the feasibility of a WBE as the cognitive experience that matters when prioritizing aspects of the brain - and not the underlying neural structure and function of which we are not directly aware. (Koene, 2018). “By analogy, internet browsers operate at the level of code and design patterns, not transistors and electrical properties. Brain emulation is, in this way, analogous to the emulation of a web browser across varying computing architectures. It all comes down to structural, but more importantly, functional, modeling of the brain by means comparable to nascent neural prostheses. Such devices, both current and futuristic, operate by modeling, replicating, and ultimately replacing corresponding neural function, one brain region at a time, until in the limit none of the organic structure remains.” (in Keith Wiley’s summary on the talk held by Koene @ the carboncopies workshop 2018).

6 _ Analogical Linkage (using paradigms toward WBE)

§1 _ Biomimicry towards AGI (and WBE)

As so often has been the case throughout human inventive and techno history, most of the grand projects began with the ambitious objective of reaching some point, scale or paradigm because of the eventual revenue it would quickly render, without first looking deep into it. Watching birds definitely gave us the idea of flying but not studying them thoroughly did make us lose ground in aeronautics engineering right at the beginning. We actually do not need wings to fly. Had we known that at the start we might have compressed quite some years of progress. Basically, we copied the concept. We didn’t copy the mechanics. Today it is clear to us we could have avoided misfortunate attempts had we undergone a deep course in biology before setting tasks and so we take it ever more seriously to copy nature in all its parallels. Biomimicry, or better yet the lack of it, could well turn out to be the reason why AI still revolves over a hazardous path in software development, keeping estimates of achieving its full potential farther beyond hardware and not showing the same pace as hardware does. “Creating the software for real singularity-level computer intelligence will require fundamental scientific progress beyond where we are today” (Paul Allen, interview for MIT review, October 12, 2011).

Alan Turing came up with the concept of computing early in the beginning of the twentieth century with a paper called *On Computable Numbers with an application to the Entscheidungsproblem* (decision-making problem). This was in 1936 during his PhD thesis and a few years later he began working with I.J. Good in the War effort to break down the German Enigma code. So, IBM built the ENIAC, which as the World's first computer, by the end of the WW2, actually proved to be critical in winning the war itself, as did Turing's effort too. In 1939, Turing developed an electromechanical device that could detect the settings for ENIGMA, a method called "the bombe," which allowed the Allies to decipher the German encryptions.

Basically, we began building computers in order to facilitate those things hard and dull for us, such as calculus. So, computers starting as efficient calculators developed into these Petascale multi-task processing machines known today and growth in computing made its path towards something frightening to us and it didn't even really happen yet. Nevertheless, though extreme progress definitely patented and whichever future AI dominant scenarios already considered today, there will always remain some kind of suspicion towards AI's eventual consciousness and in Alan Turing's own words "I propose to consider the question, 'Can machines think? A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.'" (Turing, 1950).

A computer program called E. Goostman, which simulates a 13-year-old Ukrainian boy, is said to have passed the Turing test at the University of Reading in June 2014. No computer has passed the test before under these conditions. "No one knows just why or how brain processes give rise to consciousness. Neuroscience is gradually discovering various neural correlates of consciousness, but this research program largely takes the existence of consciousness for granted. There is nothing even approaching an orthodox theory of why there is consciousness in the first place. Correspondingly, there is nothing even approaching an orthodox theory of what sorts of systems can be conscious and what systems cannot be." (Chalmers, 2010, p35).

I address this consciousness conundrum, the so-called consciousness hard problem in the mind uploading state of the art debate, in the next chapter (MU-Part III), of this thesis.

Turing's machine, as it is called - the concept of computing - resembles the biological machinery in building things, as the ribosome does. This is recognizable at a data

processing concept level. “It is also suggested by the progress of science and technology more generally: we are gradually increasing our understanding of biological machines and increasing our capacity to simulate them, and there do not seem to be limits to progress here.” (Chalmers, 2010, p8). Most of our inventions come from watching, observing and understanding nature, which therefore means to break it down, hence copying its concepts. This is of course also true in computer science, generally speaking. But just as the example given starting this particular link, sometimes we start copying nature’s concepts later rather than sooner.

In the near future, conventional computing and hardware won’t be able to cope with the energetic need to power our humankind’s incrementing massive dataset. Machine learning and AI generally speaking, have been gaining momentum recently, not only because of their natural growth rate but also because of the use of nanotechnology-based materials and consequent efficient device development. Many of them are brain-inspired approaches as some promising state-of-the-art examples herein presented.

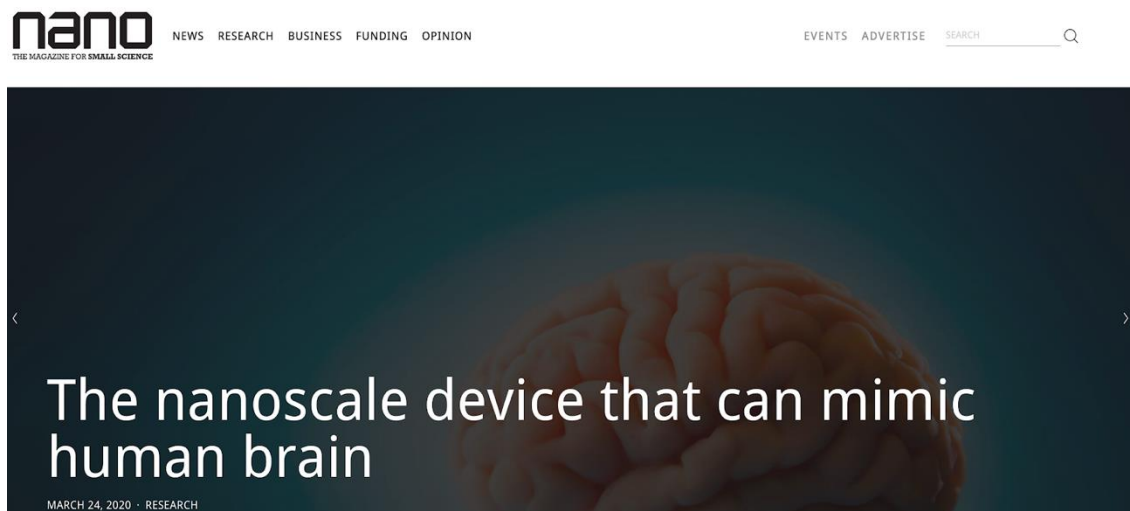


Figure 07 – Screenshot taken from cover of nano-magazine issue March 2020.

In this paper published in Nature Nanotechnology researchers of the NNI and the NUS teams have been developing circuits that mimic functions of the human brain, and invented a nano-device that can achieve optimal digital in-memory computing while being extremely energy efficient, which is something that physicists have been trying to do for more than 50 years. The device can operate both as a memristor and memcapacitor at room temperature and the team claims they can drive the devices to self-oscillate or even exhibit a purely unstable chaotic regime which is very close to

replicating how our human brain functions. Computer scientists are now acknowledging our human brain as the most intelligent and energy efficient computing system there is and achieving an emulation of the brain's performance will ultimately also change computing as we know it. (Nature Nanotechnology, March 2020). "Fundamentally, the devices we are demonstrating are a million times more power efficient than what exists today" (Venkatesen, National University of Singapore (NUS), in Nature Nanotechnology, March 2020).

Nanowire networks are a strong bet in mimicking the brain's plasticity in the sense that their electron flow changes their physical configuration as opposed to conventional electronic circuits.

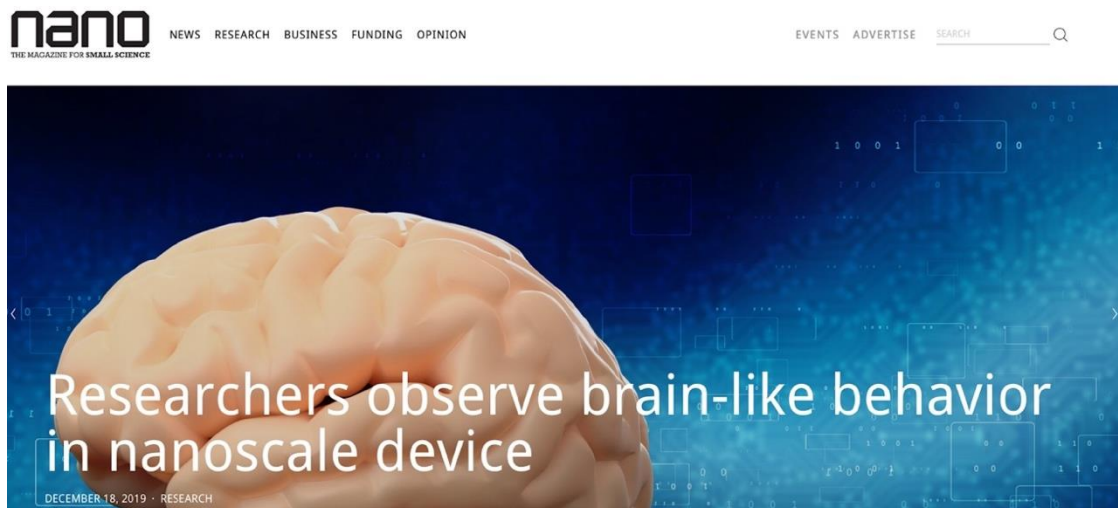


Figure 08 – Screenshot taken from cover of nano-magazine issue December 2019.

An UCLA team of researchers from the California NanoSystems Institute along with Japan's National Institute for materials Science has been developing a network kind of device in a state of continuous flux, exhibiting human brain-like behavior. "The way that the device constantly evolves, and shifts mimics the human brain. It can come up with different types of behavior patterns that don't repeat themselves." (Gimzewski, Scientific Reports 9, 14920, September 2019).

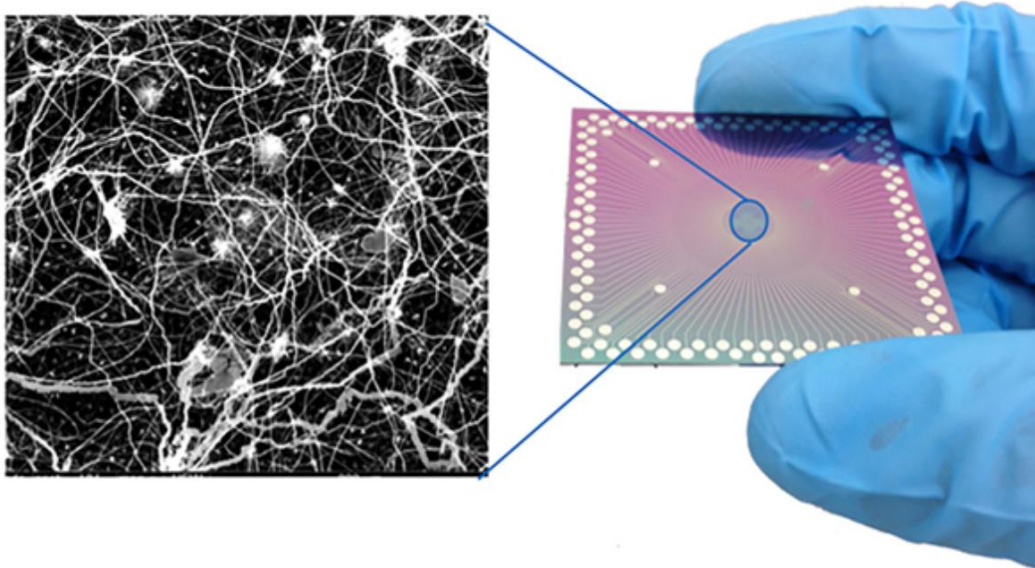


Figure 09 – A device like the one in the study (right), and an electron microscope image showing the device's neuron-like arrangement of nanowires. (Source: Photo composite by Marc Roseboro/CNSI at UCLA)

It is made out of a tangle of nanowires measuring about 10mm². These nanowires randomly self-assemble on a silicon wafer forming interconnected structures in a similar way the neocortex does in human brains. Connections between overlapping nanowires are observed whenever the current flows through the device comprising about 10 million, of these junctions (analogous to synapses). Another promising characteristic observed was the fact that once electrodes were attached to the device in order to study its networks' overall performance, a kind of emergent behavior was reported as something working as a whole as opposed to an individual behavior of the components which, again, is a similar behavior to the human brain and quite different than the one reported in conventional computing. The study even reports resemblance in learning and memorization processes in the sense that connections between nanowires persist for some time after the current flows through the network whereas turning the current flow off also shuts down any connections, somehow like the human brain's forgetting process.

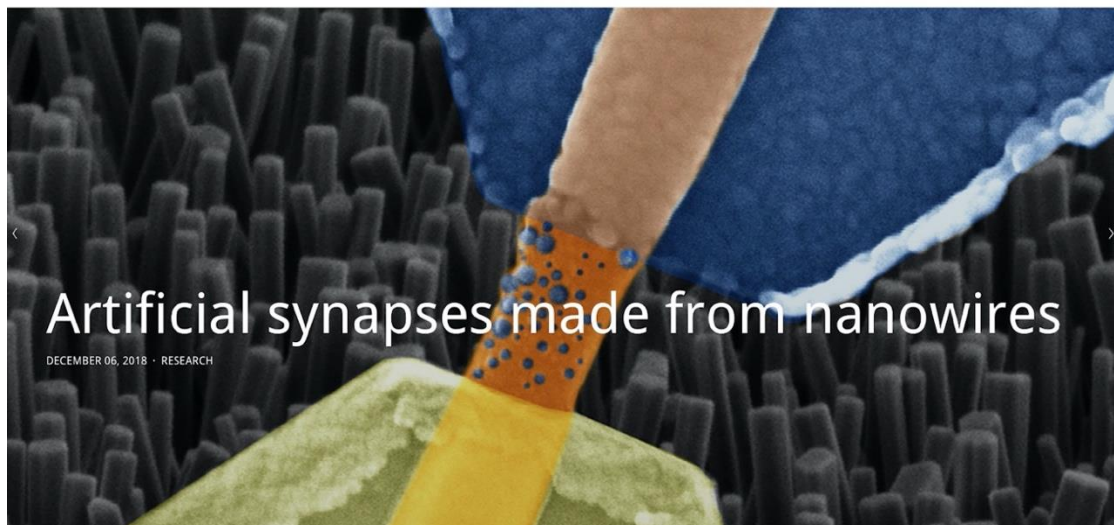


Figure 10 – Screenshot taken from cover of nano-magazine issue December 2018.

Memristive cells, as described in the previous section's examples, have been one of the promising approaches in mimicking neurons and synapses in computing since they can process but also store information in a similar way, neurons do. As opposed to conventional computing, where traditional transistors do not exhibit a 'learning capacity', these nanowire devices combining memristive and neuromorphic functionalities, keep the latest resistance values even after switching off. (Gianluca Milano et al, Nature Communications, 2018). Moreover, they are able to process in parallel, which again, is something traditional transistors cannot do but the human brain does. This is actually another big challenge. Neural networks process information in parallel whereas classic computing works in succession, dividing tasks, which makes it inefficient to simulate neural networks on conventional hardware. Neuromorphic chips work in a de-centralized way using multiple processors, and like the brain does with neurons and synapses, are connected to each other by networks, thus imitating nature where processors once broken can be substituted by others with the same function. Ultimately the goal is to improve as to make a bioinspired processor to learn which can be done with these nanowire devices that inherently process and store information energy-efficiently. (Valov, I., 2019).

Devices based on nanowire technology have been shown to be energy-efficient, as the human brain also is, which is quite different from the tremendous work power needed for

conventional computer servers, that whether for machine learning or to internet searches, consume households' amounts of energy.

Furthermore, these nanowire networks are brain-inspired approaches, both energy-efficient and capable of processing complex datasets as opposed to conventional computing hardware.

§2 _ Exponential Progress Rate

If we relate this cost-effectivity with the metrics of cps/\$ (calculations per second per dollar), as used by Kurzweil (plot in figure below), then we'd be around 10 trillion cps/\$1,000 by now (around 2015), and thereby inside the expected boundaries of the ratio. Computers of today of the \$1,000 order, rival with mice brain capacity, still one thousand times off the human level. But by 1985 they were a trillion times off, by 1995 a billion and by 2005 around a million. Being ca 2015 around a thousand times away could mean that at this pace we could expect computing power to be sufficiently powerful to make some impression on the human brain around the year of 2025.

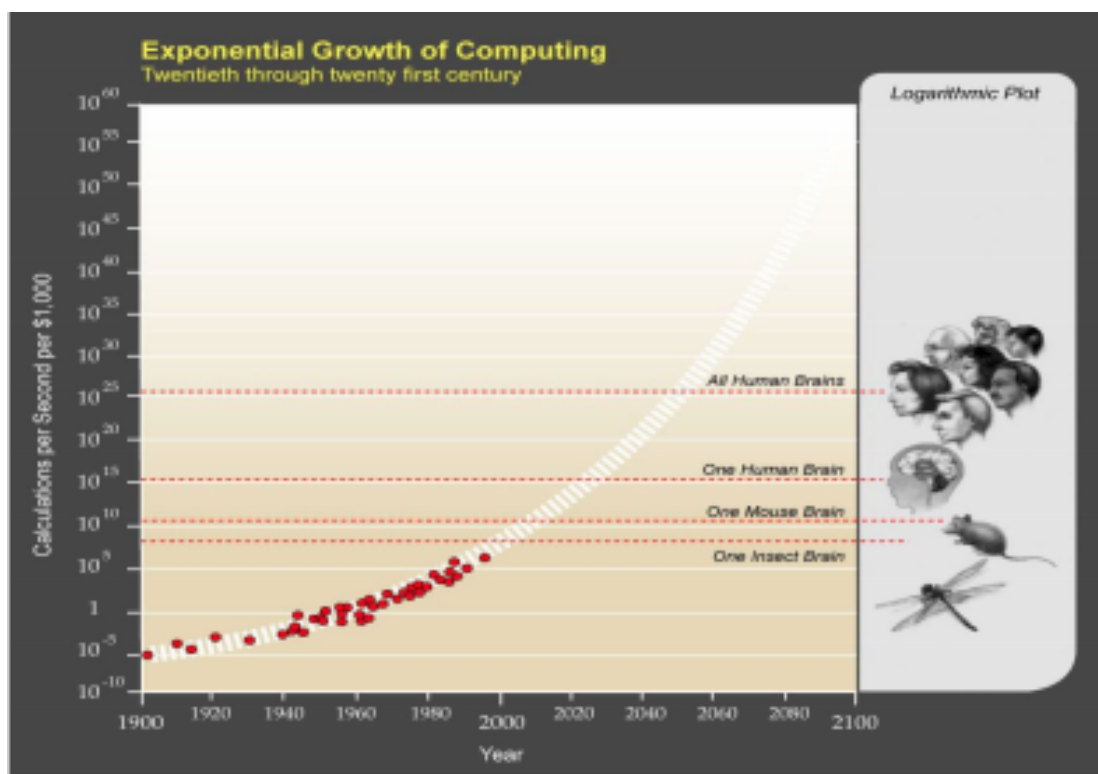


Figure 11 – Exponential Computing Growth measured in CPS/\$ following Moore's Law. (Source: Kurzweil, 2005)

Considering Kurzweil's plot for the cost-efficiency metric used to compare biological brains to the growth of computing and artificial brains, then we have to take into account that the actual ratio we are in holds true for the trend and achieving an AI as powerful as the human brain in the next decades is not sci-fi at all. And knowing that Kurzweil is actually currently leading AI projects at Google it seems only fair to state: "A Google-mapped brain is underway, and it could not only lead us to understand and cure diseases, but it could indeed end in a map of consciousness. Just as today's sophisticated circuit-editing tools can modify microchips after they've been manufactured and packaged, a brain-editing tool could perhaps one day modify the brain" (Adler, 2012). Exploring the Brain using computer technology to integrate everything providing data about everything on a molecular level from genes until cells, their production, expressions and connections is a quite recent enterprise in our history so it should not shock us that the Brain itself remains our greatest mystery and decoding it eventually our grand achievement.

Starting around 2013/2014 the HBP (Human Brain Project) entails a number of research studies at different levels of organization and development and aims to upload the entire mouse brain (workflow overview of the Whole Mouse Brain Modelling*) in a decade and upload parts of the human brain in the following years. Co-funded by the European Union it had set high ambitions. However, its main target to unravel the replication of the human brain has up to date failed according to the initial timeline established. Other projects such as the Blue Brain Project are also all relatively recent and striking results will take some time but usually the take-off in an exponential ratio starts with a slope linear stuttering kind of way. The Human Genome Project, the complete mapping all the genes of human beings, began in 1990 and took roughly 15 years to complete. But half-way of it still only less than 10% of them were covered so people in general were given the impression it would take an eternity to decode the human genome. It was only in those last years (the full sequence was completed and published in April 2003) that a number of features combined together triggered off and the project made its real leap forward. Comparisons are a dangerous and complex task and shouldn't be done lightly but this could explain some of the possible under-achievements or near misses. Nevertheless, what they do tell us and the lesson we can take from here is that disciplines are combined to form synergies and knowledge put together, applied to solving a specific problem, establishing a roadmap and financially holding on to it, this endeavor is almost as exponential as progress itself.

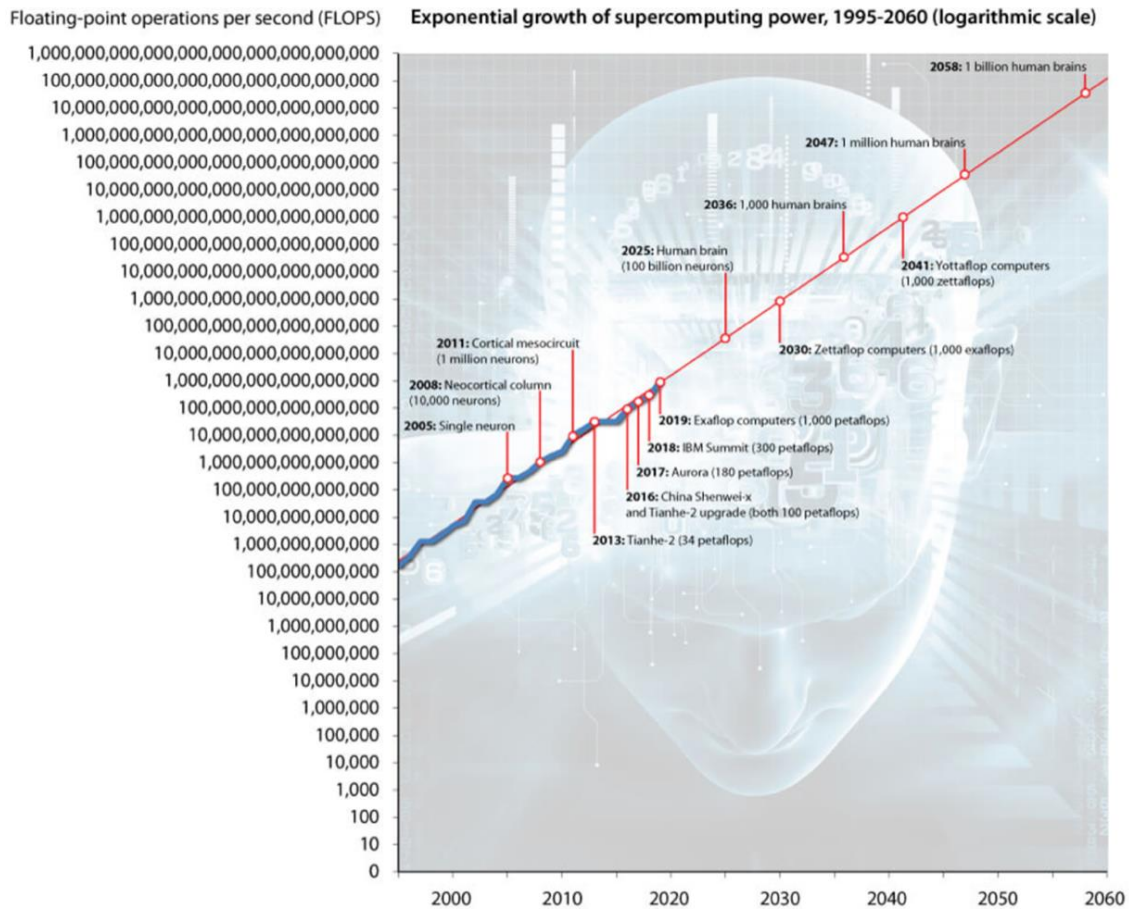


Figure 12 – Exponential growth of supercomputing showing estimated timeline of reaching and surpassing human brain capacity in FLOPS. (Source: available on the internet)

§3 Classical - Quantum Transition

Moore's Law goes beyond transistors or specific technologies. It may be a paradigm related to humankind's creativity. "It appears that what Moore really uncovered was humankind's ability to cost-effectively accelerate technology performance." (Del Monte, 2013, p28). But Moore's law won't last forever, and the physical miniaturization limit of transistors will naturally condition that trend sooner or later. Nevertheless, Moore's silver-bullet prevails, for it is more a concept than simply an observation, or even a law. It will duly be replaced by the next paradigmatic change. Technologies will always replace each other in time and in progress naturally. Actual trends and curves prove it to be true and so do their extrapolations. And new technologies are emerging to use molecules individually positioned, replacing transistors altogether. This means computer *switches*

will replace transistors for molecules. The position of the molecules will be the new switches. This technology is predicted to emerge by 2020. (Waldner, 2008).

The new computer technology following Moore's Law may be based on a new type of technology (e.g., optical computers, quantum computers and DNA computing) that bears little to no resemblance to current integrated-circuit technology.

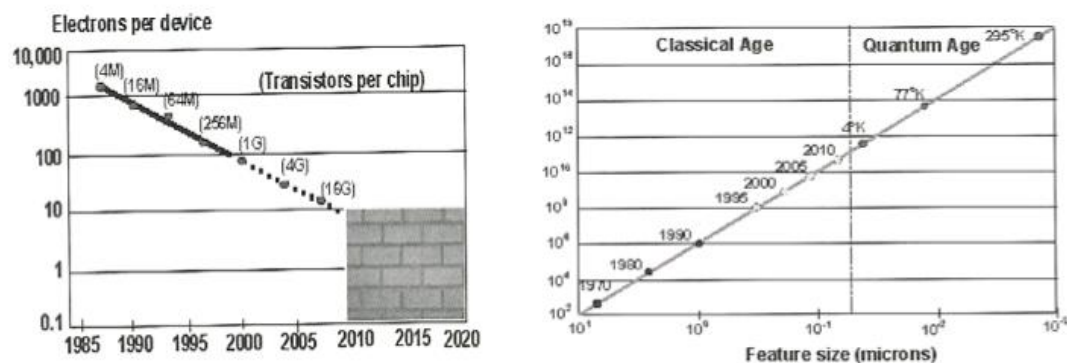


Figure 13 – On the left physical miniaturization of silicon devices is hitting the wall of its limit, which is coincident with the transition of the classical to the quantum age (on the right). (Source: Wilson, 2002).

As we can infer from the above shown plots: On the left, physical miniaturization of components reaching its limit with silicon transistors coincides with the entrance in the quantum age shown on the right plot. This may well be the paradigm shift enabling us to leave classical computing in favor of quantum computing which in turn would also mean that Moore's law would still be verifiable. The upcoming years of the 2020 decade will prove this to be true or not.

We could be witnessing this paradigm shift as it happens. From a classical physics dependent kind of world still operating today towards a quantum related one. The new era we're delving into is also going to bring about a shift in the way we think. Our minds are hard-wired to a linear thinking kind of way. But with the abovementioned paradigm shift they might become more exponential in time, which of course is a speculation.

Back to that analogy right at the beginning of this chapter, linking biomimicry (or the absence of it) to the software underachievement in progress and to Reverse Engineering the Brain: wouldn't it make sense to assume that, had we approached computing before getting to AI (weak or strong AI, that doesn't make a difference at this stage) using Nature

and its products to copy by as a reference guide and software could have the same pace as hardware does by now?

Exascale computers are expected to deliver cellular level simulations of the complete human brain with dynamic switching to molecular-level simulation of parts of the brain when required in the near future as mentioned earlier. Thus, this reverse engineering, will help us explore and predict how different patterns of gene expression produce neurons with different morphologies expressing different molecules, and different synaptic connections without resorting to invasive methods of data collection. (Markram, 2011). Neuroscientists combined with computer scientists have been developing tools and algorithms, artificial multi-level brain-like neural networks similar to the ones measured in the brain itself. We've been reshaping technologies from pattern recognition to autonomous driving using algorithms inspired in the way our brains work. Human technological capability is in a deep learning revolution and it's boosting a new Era in AI.

Superposition of states allows taking advantage of all the space that goes from 0 to 1, as opposed to binary states, as seen in Figure 15 (below). The probabilistic waves enable to store and process an infinitude of data when compared to classical computing, which could mean that quantum computing could be the next paradigm keeping Moore's trend alive, as mentioned earlier, but also tackling the lack of sufficient hardware and software necessary to process brain data.

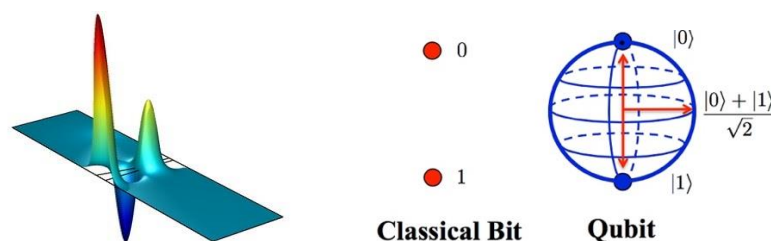


Figure 14 – On the left representation of classical Bit. On the right the quantum correspondent - Qubit. (Source: available on the internet).

§4 _ Synergy (and connectomics)

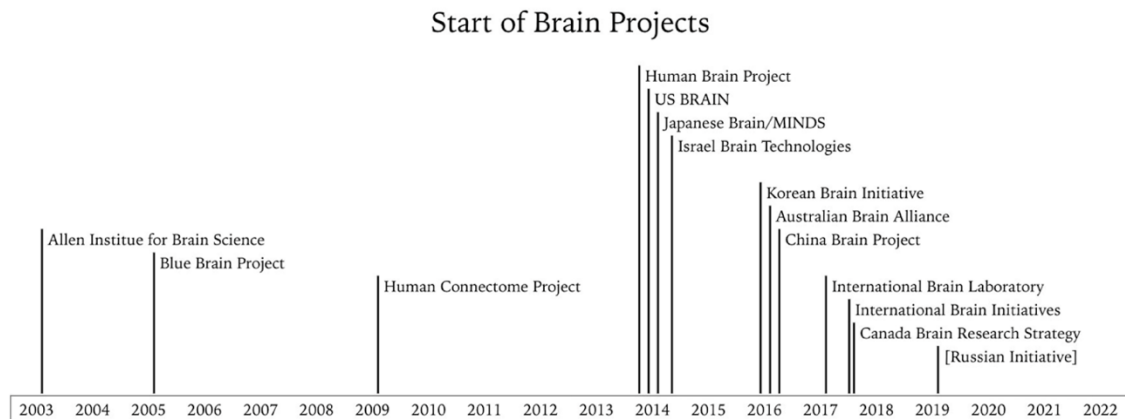


Figure 15 – Brain Projects Timeline. (Source: available on the internet)

There are still strong issues haunting a possible AGI when compared to the human brain capabilities. Supercomputers have thousands of processors, but the human cortex has billions of neurons and trillions of synapses which themselves vary widely in cortical neurons turning cortex emulation into a big wiring problem. Besides these neural complexity and scalability issues there is also the plasticity issue. It is clear that synapses are plastic with connections changing with new information.

Furthermore, the human brain evolved to consume as little as an estimated 25 watts whereas an artificial setup to emulate the brain with such an efficiency would need to generate another magnitude of order, although nanotechnologies seem to promise a lot also in this department. (Cattell & Parker, 2012).

To overcome these issues various projects besides the HBP with different approaches are being led. In the short-term software simulation seems promising for emulation of small networks of neurons whereas in the long term neuromorphic analog integrated circuits with nanotechnologies and the very promising self-assembly* concept should play a major role. Self-assembly could solve the neural-interconnection issue since it basically delivers cell orientation which would self-guide the wiring of dendrites and axons. (Cattell & Parker, 2012).

The Blue Brain Project started a little over 10 years ago, with the BlueGene Supercomputer from IBM, has already made significant advances in building digital reconstructions of the brain at an unprecedented level of biological detail. The BBP uses

its digital reconstructions for simulations represented in silico experiments. It also allows to measure electrical activity of the virtual tissue, apply stimulation protocols and measure the responses over complex interactions within different levels of brain organization and to investigate the cross-level links leading from genes to cognition.

On the other hand, the plasticity issue – the way neurons learn over time through morphological and functional changes in synaptic connections – is of course also being thoroughly explored by other projects such as is the case for SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics). The project developed nanoelectronic components capable of adapting connection strength just as happens to neural plasticity in biological brains. (Deca, 2011) This will, through reverse engineering, also help us to better understand the neural learning ability since it is still quite unknown. There is research indicating that glia (Fields, 2013), astrocytes in particular, communicate and perform an active but not thoroughly understood role in network activation and plasticity.

Recent developments in nanoscale analysis tools and design of novel nanomaterials have generated electrical, chemical and optical methods that are being adapted to be used in neuroscience and new developments and advances are starting to unravel many of these aforementioned issues on an unprecedented timescale. Some of these state-of-the-art examples are resumed below:

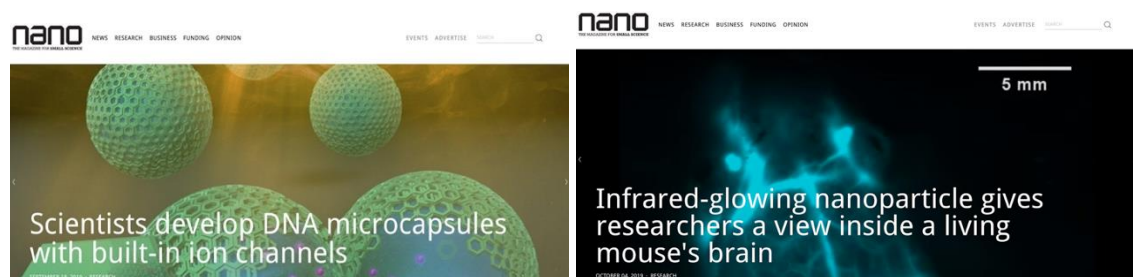


Figure 16 – Screenshot taken from cover of nano-magazine issues September + October 2019.

Researchers of the Tokyo Institute of Technology have been developing stimuli-responsive systems that in turn could be used to develop artificial neural networks similar to the human brain. This approach uses DNA-based self-assembled nanostructures that could mimic the functionality of ion channels that the brain uses in its synaptic connections. (Takinoue, 2019). (on the left in the figure above). A new deep-tissue imaging technique that non-invasively peers through biological tissue has been

developed in Stanford (on the right). The team of chemists of the Stanford's School of Humanities and Sciences used an engineered coating that dissolves the nanoparticles in the bloodstream and makes the target tissue cells glow under a low-powered LED light. It allows for an unprecedented spatial and temporal resolution and enables the possibility to better target cells in a non-invasive way with reports also claiming a great molecular specificity than with conventional imaging techniques so far. (Dai, Jackson and Wood, 2019).

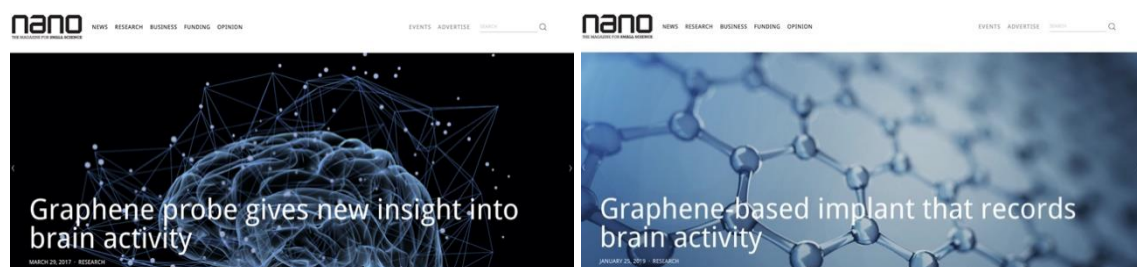


Figure 17 – Screenshot taken from cover of nano-magazine issues March 2017 + January 2019.

European Union's Graphene Flagship has been through the joint-venture kind of synergetic work assembling different teams from Germany, Spain, (and more) developing nanodevices to record brain activity in high resolution. "The flexible array of graphene SGFETs allows mapping brain electrical activity with excellent signal-to-noise ratio (SNR), suggesting that this technology could lay the foundation for a future generation of in vivo recording implants." (Blaschke, 2017, p1). The researchers hope to take these preliminary epilepsy related studies in rats to the proof-of-concept level in human clinical trials. These graphene-based devices pave the way not only in brain mapping potential but also in the development of brain-computer interfaces. By amplifying the brain's signals in situ before transmitting them to a receiver they can record at very low-level frequencies. The flexibility of the implant, as well as it being a slim device, allows it to be used in the cortex without interfering with the normal brain function. As a result, it is possible to map brain activity of low frequency that is known to be tied to critical information about different events. This device was developed by various Graphene Flagship Partners (IMB-CNM, ICN2 and ICFO). (Nature materials, 2019). This research coupled with other techniques, including neural net simulations can lead to an understanding of various neural subsystems. This seems to fuel optimistic predictions, as Smith claims: "One reason for our considerably more optimistic predictions is that brain emulation relies on computing and other information

technologies, fields that have classically enjoyed exponential gains in capacity and should continue to do so for the foreseeable future. Computational neuroscience is now big science, and significant resources, above and beyond previous investments and commitments, are now being devoted to emulation. See for example the UC-funded Human Brain Project. Basic neuroscience tools and funding have emerged for mapping and understanding the connectome and the synaptome.” (Cerullo, Why Brain Emulation is coming sooner than many think - a bpf response to dr.Miller’s NYT editorial on brain emulation, October 27, 2015).

Finally, as a highlight of synergy (and connectomics), The Carboncopies Foundation, a non-profit organization, provides support to scientists working on challenges critical to WBE as stated in their mission’s message on the official site. Their chairman (Randal Koene) recently announced in a workshop held at Carboncopies, the launch of a tech review project with the goal of maintaining a living roadmap of projects and progress. This tech review will seek peer-reviewed research into the technical aspects of WBE. (Webinar: Transcending biology: reverse engineering the brain, Feb 2018).

§5 _ Brain Preservation (and postponing the reset)

Cryopreservation agents significantly damage the brain at the molecular level though much progress has been done in post-mortem brain preservation in recent years. Alzheimer’s disease, ischemic seizures or strokes irreversibly damage the brain making a person’s connectome recovery impossible. Even the preservation process involving crystallization unable to restore the brain to the size of a human’s brain, though it has been accomplished with smaller ones.

Kenneth J. Hayworth, co-founder and president of the Brain Preservation Foundation (BPF) whose goal is to seek and promote long-term whole brain emulation, emphasized the need for specific techniques for SEM scanning of brain slices for WBE (Hayworth, 2012). According to McIntyre, expansion microscopy aldehyde-stabilized cryopreservation enables indefinite storage of whole brains while preserving the ultrastructure with minimal distortion and scalable to any sized brain. (McIntyre, 2015). If neural wall structures are intact, dendrites, neural cell types, soma configuration and axons can be traced along with the synaptic junctions then one can reconstruct a

network. Also, it may be possible to ultimately reconstitute a person's memory and personality as soon as mapping data concerning functionality is reliable. This in turn might be achieved when key information concerning the synaptic strength and the neurotransmitter type has been successfully identified. (Feierbach and Koene, 2015). There was a challenge launched by the BPF to develop a reliable brain preservation technique that would enable to preserve the synaptic connectivity of a post-mortem human brain in a way to be stored for over 100 years. As a result, the cryobiology research lab led by McIntyre, 21st Century Medicine (21CM), uncovered limitations in the current cryonics protocols when judged against the goal of verifiably preserving synaptic structure. 21CM invented a new brain preservation protocol, called “Aldehyde Stabilized Cryopreservation”, which appears to have entirely overcome these limitations. (brainpreservation.org).

The BPF aims for brain preservation to become a standard medical practice in hospitals for life-preserving purposes. Hayworth believes that brain preservation could be used to bridge the intervening timespan until mind uploading would become feasible. (Hayworth @ webinar “WBE - Roadmap 2018, Where are we now? 2018). The speculative timeline presented below (author Ken Hayworth) was presented in the aforementioned workshop and indicates an elusive fictioned neuroscientist (called Joe) who does not live until mind uploading could be achieved. But brain preservation could tackle this issue if it indeed would be achieved earlier (as indicated in the schematic), and Joe’s mind could be then uploaded, *a posteriori*.

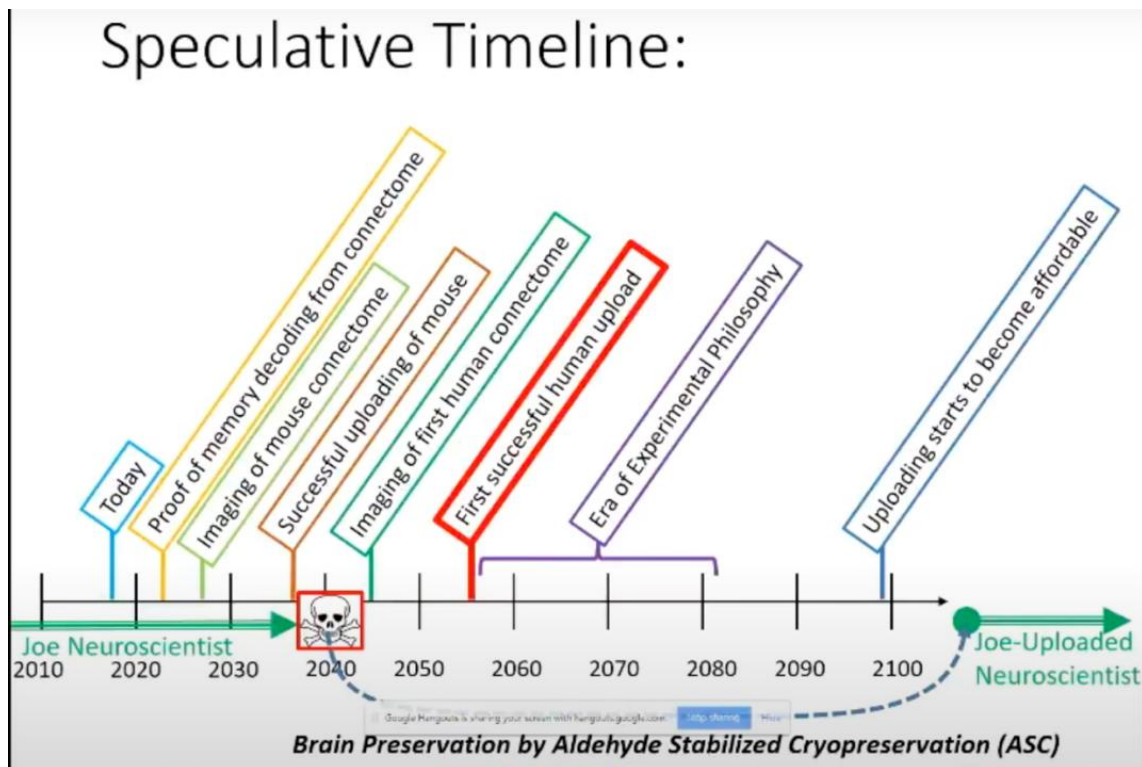


Figure 18 – Screenshot taken Speculative Timeline during Hayworth’s Talk. (Source: Hayworth, 2018).

7 _ Preliminary Concluding Remarks B

Critics of WBE frequently argue that a WBE would be incomputable because features of consciousness could be the result of unpredictable nonlinear interactions among billions of cells. But I would argue that being unpredictable doesn't make it unfeasible. Again, the fact that we don't understand these interactions does not mean we can't copy them. Ultimately it depends on the validation criteria chosen. As Koene remembers, “We can have Mac emulators run on PCs and we sure do not know their atomic level precision on either of them.” (Koene, @ the carboncopies workshop 2018). Also, how can we know what are features of consciousness, without even knowing what consciousness actually is?

With increased computational power (which is going to accelerate even more as soon as operating on carbon nanotubes) sufficient knowledge of synaptic and dendritic

connectivity should be achieved allowing for the complete emulation of brain function. (Cerullo, 2014).

Neural prostheses, WBE and mind uploading are in this order of increasing complexity. There is already today a large number of neural prosthetics substituting, improving or augmenting specific sensory functions (cochlear implants / retinal implants / electrodes to inhibit seizures). These prosthetics are developing towards a hippocampus prosthesis that would then later on eventually enable a WBE on a non-biological substrate.

Theodore Berger (University of Southern California) presented the state-of-the-art in neural prosthetics at the aforementioned workshop (2018). “Berger’s prosthesis applies electrodes to two regions of the hippocampus, loosely labeled as the input and output areas. It learns the neural spatio-temporal firing code from which it is possible to learn the transformation function performed by these regions of the hippocampus and can actually take over or strengthen this transformation for a damaged hippocampus. In this way, the prosthesis revitalizes a patient’s ability to form new long-term memories (...) Taking this concept to completion by replacing every brain region with an equivalent prosthesis is, in fact, whole brain emulation.” (in Keith Wiley’s summary @ the carboncopies workshop 2018).

In this thesis, it was my deliberate intention to present several concepts relating to different techniques, all of them analogically related to one another and to WBE, such as biomimicry (§1). This seems to be done at the right time since most of these brain projects are all quite recent and along with the exponential growth (§2) in computing they too could become ever more exponential. Not only because communicating exchange of valuable technological info (synergy & connectomics) (§4) of course accelerates the whole process but also because we might be at that slope of the plot just before lift-off. At least trends and estimates point in that direction.

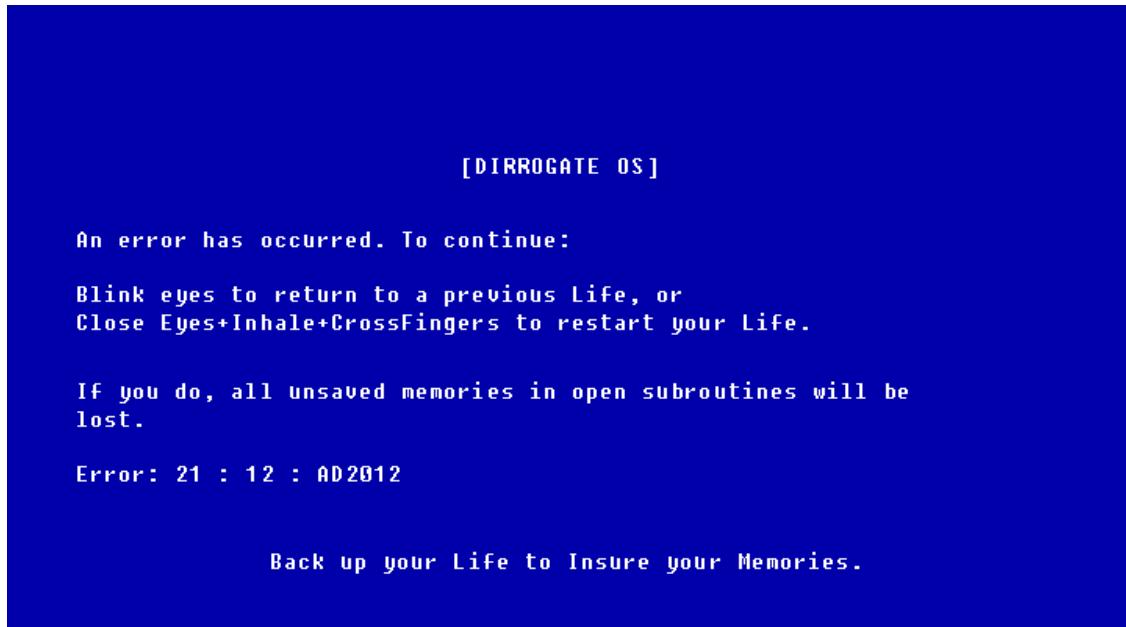
Furthermore, as already claimed in this thesis, nanotechnology-driven tools are already starting to display some groundbreaking applications here shown in this section as state-of-the-art examples underlined through some of their most recent developments. And as soon as we trade Newtonian physics for the quantum one, once silicon miniaturization and Moore’s Law have dried out, those tools will once more quantum leap (§3). Might we not by then have sufficient technology for WBE?

We live in a time in which even economics is not only technology-driven but technology-dependent, and universal Darwinism seems to prove itself to be true every second goes by, pushing competition to excel innovation faster and faster which surely is not going to ease sceptics arguments. However, as stated in analysis of the interviews made for the purposes of this thesis, historically the trend always seems to be on technology's side and breakthroughs and thresholds are just around the corner.

The EM (electron-microscopy) preservation of a whole rabbit brain (achieved by McIntyre et al at 21st Century Medicine, 2016), a peer-reviewed and verified feature proving high-quality preservation of a whole brain not only represents a milestone in brain preservation [§5](#) but also a proof-of-concept toward a WBE. Also, Kenneth Hayworth's pioneering developments of slicing and imaging machines, in which a stasis-preserved brain's structural scan is used as the basis for a subsequent WBE, constitutes a breakthrough in that critical step (even critical issue) of brain scanning, hence a landmark toward that goal. (Webinar: Transcending biology: reverse engineering the brain, Feb 2018).

Stated as some sort of slogan, and after most recent roadmap updates, we may say that WBE may be trending to be the technology of "preserving, scanning, imaging, mapping and functionally reproducing the brain." (in Keith Wiley's summary @ the carboncopies workshop 2018).

PART III _ Mind Uploading



AI makes philosophy honest.

— Daniel Dennett (2006)

1 _ Intro / Technical background

Mind uploading might be considered to be the logical endpoint of ongoing whole brain emulation projects (such as the HBP and the BRAIN Initiative, these two being the most prominent, at least financially and given that they are government-sponsored) as outlined in the previous WBE chapter. However, none of these projects are specifically aiming for a mind uploading. They are rather aiming for emulating the information processor machine called the brain. But this eventually also meaning *the mind*. This raises a lot of issues, such as those of consciousness or identity, as I shall discuss in this chapter. However, when discussing the concept of a mind uploading, one question that is usually immediately put forward is the one concerning its feasibility and timeline.

And concerning its possible timeline, one cannot overlook the predictions made by Ray Kurzweil. And concerning these forecasts, we should start by noting that in spite of Kurzweil's successful predictions rate (about 94% of these,, Kurzweil claims are essentially correct, and 86% reasonably correct predictions, Kurzweil, 2012), fulfilling the expectation of achieving a successful mind upload by the 2030s does seem unlikely and too optimistic . A broader timeline, like the one projected by Kenneth Hayworth (shown at the end of the WBE chapter), with an open window range of about 50 years is a more conservative estimate and, as such, perhaps we can argue that it is doable. One conclusion stemming from all the uncertainties outlined in the previous chapter for WBE, an epistemically reasonable projection would have to be conservative. Nick Bostrom also concurs with this assessment: “the prerequisite capabilities might be available around mid-century, though with a large uncertainty interval” (Bostrom, 2014).

Following Sandberg's roadmap of 2008, and Koene's updates of 2012, as well as Koene's more recent ones of 2018-2020, whole brain emulation's feasibility could be roughly divided in three (3) essential *key moments*:

1) Scanning: The neurological information of the brain can be obtained by scanning it and performing a sort of Data Collection. This acquisition could be addressed through high-resolution microscopy technologies, but as stressed out in the WBE chapter, these are invasive and therefore destructive. This brain imaging together with other ancillary technologies would acquire the functional properties of neuronal interconnections, synapses and ion-channels, and thereby capture the communication processes involved

inside and between different parts of the brain, enabled its mapping. Hence through reverse engineering we could understand the brain better. As such, WBE would allow to map the exact brain functionality on a one-to-one mapping basis. Yet, as stated, there is still uncertainty regarding the level of resolution needed to register and detect all the necessary details for a simulation to take place, which makes this issue very much dependent on the continuing development of those technologies and techniques.

As already mentioned in this thesis, nanotechnologies of course hold a big promise in this department. Not only because of the simple fact that they define atomic manipulation, therefore sooner or later eliciting nano-probes that could record or even reproduce neural activity, as some of the state-of-the-art recent developmental examples shown in the previous chapter, but also in developing the tools to improve scanning itself.

Nanoprobes in the speculative form of nanobots (Drexler, 1986, Freitas, 1999 and Kurzweil, 2005) would certainly stand as an ideal method, but its ideality is somehow directly proportional to its speculative status. Replacing every single piece of the brain through electronic devices maintaining functionality to then interface with the rest of the brain is actually an idea pioneered in science fiction literature in the 1960s and 1970s (and even earlier) by acknowledged sci-fi authors but that also were acknowledged and laureate science-making authors, such as Isaac Asimov and Arthur Clarke (amongst others). But it was not until the late 1980s, after the first milestones in nanotechnology came through, such as the development of the STM (scanning tunneling microscope) and a bit later of the AFM (atomic force microscope) that the huge capabilities offered through this kind of matter manipulation gained popularity (Drexler, 1986, Moravec, 1988). By replacing piece by piece components of the brain (whether they would be neurons, ion-channels or else), all that would remain in the end would be an electronic device with the original's system information. These last couple of decades saw another huge leap in technological development leading many authors to speculate today that a human brain could in fact in the near future be scanned at a high level of detail and that functioning could be derived and simulated from this data at a neural level, by a computer. (Kurzweil, 2005 and 2012; Sandberg & Bostrom, 2008; Hayworth, 2010; Koene, 2012).

2) *Translating*: The collected data entailing functionality of the connectome altogether would then be translated into algorithms that could be used as the initial state for the simulation to take place.

3) *Simulating*: Simulation would be carried out from the aforementioned algorithm containing the model designed out of a network similar in functionality to the original system. Again, as stated before, this step would also require an enormous hardware capability in processing or in memory (or in both), always depending on the required detail level necessary for the emulation to occur. But as discussed earlier, this computing power needed does not seem to represent an uncertainty as big as the one regarding the scanning step.

Apart from spatial constraints at the resolution level, there are also time constraints playing a role in scanning techniques. The speed and parallelization likely needed to avoid huge scanning time has not yet been achieved with current technologies available. Nevertheless, another promise relies on brain preservation techniques, as approached in the previous chapter. These would allow for serial sectioning of frozen brain tissues and enable analysis of slices, layer by layer, that can capture neuronal structure and their interconnections. So although being a destructive technique to cut slices of the brain, preservation techniques have shown to be capable of bypassing this issue with some of the groundbreaking examples also mentioned in the previous chapter of this thesis. These advancements seem to point in the direction of being able to permanently preserve the complete human brain in the coming few years which would then enable us to store it for a very long time without loss of information until uploading would become possible. (Knott et al, 2008; Hayworth, 2012).

If scanning and the following emulation would succeed, the upload could be then embedded in a kind of virtual reality environment or even robotically embodied. But these envisaged possible scenarios (addressed further on toward the end of this chapter in the sections 6.3 and 6.4) definitely seem less worrying in terms of feasibility than any other of the pre-upload steps mentioned above. Information collected about an individual could be used to parametrize a generic substrate (Rothblatt, 2012; Bainbridge, 2012) before reconstructing an individual, as well as recording the low-level functioning of the living brain as a complementary data-gathering approach in order to assist in the parametrization of such simulations, which is also another proposal to adequately address post-uploading issues. (Koene, 2012; Martins, Erhlagen and Freitas, 2012).

In the near future, we could be facing brain preservation options (and their following choices) at (or right before) death, given accelerating progress in neuroscience. We would then choose not only from an individual perspective, but taking into account the

implications tied to a variety of outcomes impacting humankind at all levels. As will be discussed in the course of this chapter, it seems likely that the first uploading method to become technically available will be destructive, and based on scanning brain slices. This raises many questions concerning consciousness and personal identity that impact dramatically on philosophical, psychological, ethical, sociological, cultural and political issues. Many authors criticize the whole MU project, claiming it to be a sort of *an expensive scam* that might not be more than an assisted kind of suicide from an individual's perspective. Sceptics generally try to refute MU's technical and scientific credibility as discussed in the course of this chapter along with my tentative counter-arguments. Delving into the existential debate, it will be argued in this thread that MU may become not only possible but even essential in humankind's evolution.

2 _ Approaches to Mind Uploading

As of today there are two (2) main possible procedures described in the literature regarding Mind Uploading (henceforth referred to as MU). Truth is that given the reasonable estimates mentioned earlier (i.e. to have MU until mid century), there still is some time for technologies, generally speaking, to develop. By this I mean that the procedure that might eventually be successful in achieving a MU could be none of those now envisioned. Nevertheless, and as assessed from the best available science today, from a conceptual perspective the two main options would be: 1) a gradual uploading (also called *gradual in-place replacement*; or 2) an instant uploading (also called *destructive scan and copy*). The former is somewhat erroneously believed to be the preferable one because it supposedly holds better chances of preserving personal identity whereas the latter supposedly would produce some other identity in the target substrate such that personal identity is lost along with the biological brain. (Wiley, 2014). This metaphysical debate will be duly addressed in the personal identity section further ahead in this thread. Allow me to briefly explain these two main options.

1) Gradual Uploading / *gradual in-place replacement*: This approach could be that of a gradual replacement of individual components of the brain with devices of functional equivalence. (Wiley, 2014). If the level of information determined would require that of a

single neuron, then the replacement could be that of a single neuron, piece by piece, eventually through nanobots. Either by them recording the neuronal activity and then transmitting the information to the computer, or substrate destined to house the emulation, or by taking its place right after decoding it, i.e., after learning about it and about its connectivity and then replacing it when ready to simulate its behavior. (Chalmers, 2010). This procedure would be carried out until *every single neuron* would have been replaced enabling and ending with an emulation on the new substrate (most probably a computer).

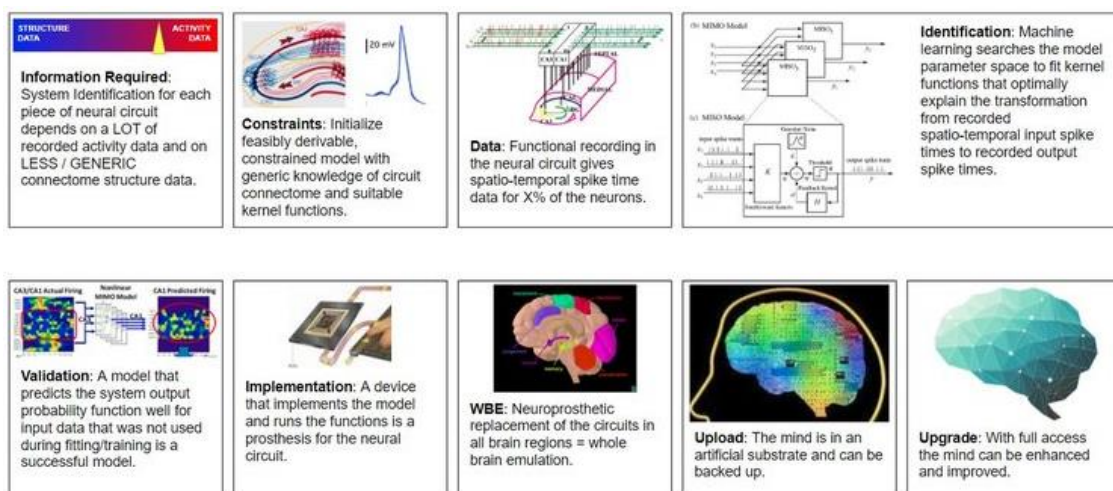


Figure 19 – Gradual replacement w/ neuroprosthetic devices for each brain piece. (Source: carboncopies 2018 / <http://www.minduploading.org/>)

2) Instant Uploading / destructive scan and copy: Since a cryopreservation of the human brain seems to be feasible in a near future, serial high resolution sectioning of layers of the brain, containing all relevant features and components entailing the connectome's functionality, would account as the most probable kind of MU to be available first. The instant scan would then be uploaded into the computer with the appropriate template or model to run an emulation of the original system (brain). (Sandberg & Bostrom, 2008; Koene & Deca, 2014; Wiley 2014)

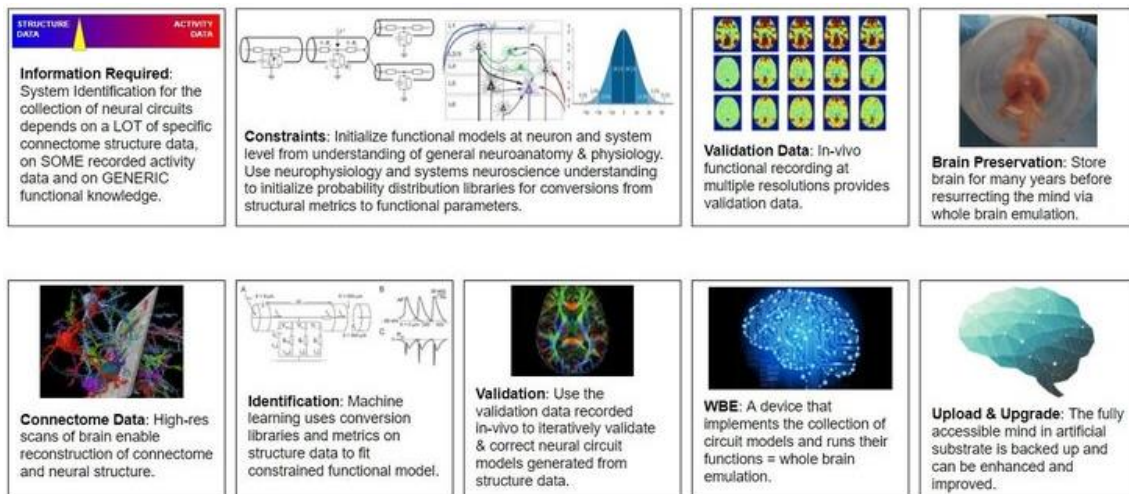


Figure 20 – Reconstruct a working artificial brain from the information in tiny brain slices. (source: carboncopies 2018 / <http://www.minduploading.org/>)

The schematics were taken from a *carboncopies* webinar held in 2018 and represent the technical steps envisaged for gradual replacement uploading and scan-and-copy uploading scenarios respectively.

Gradual Replacement is popularly considered to be preferable to scan-and-copy regarding the preservation of consciousness and particularly that of personal identity (sometimes called our ongoing experience of ourselves) (Block, 1995; Chalmers, 2010; Olson, 2010) essentially based on the argument that it could achieve personal survival whereas scan-and-copy cannot. Furthermore Gradual uploading is regarded as eventually successfully preserving identity whilst scan-and-copy would at best produce a copy or even a different person. (Corabi & Schneider, 2012; Morris, 2013; Dorrier, 2015).

In non-destructive uploading (which corresponds to the gradual replacement or gradual in-place replacement uploading scenario) a nanotransfer in some sort of noninvasive brain imaging like fMRI could be able to record neural and synaptic dynamics. But this kind of technology is of course not available currently whereas the destructive uploading (scan-and-copy) shows evidence of becoming technically feasible in a short term and is held to produce the exact same result, metaphysically speaking, i.e. preserving consciousness and personal identity. (Koene & Wiley, 2013; Wiley, 2018).

3 _ Consciousness Conundrum

“Every man’s world picture is and always remains a construct of his mind and cannot be proved to have any other existence” (Schrödinger, 1958)

What is consciousness? There is no answer to this seemingly simple question. Yet when you hit “*consciousness*” in a simple google search you get 330.000.000 results (as of June 28 2020) of which roughly one third ($\frac{1}{3}$) rely on explanations and opinionated views on the subject (About 131.000.000 results).

The first hit gives you the dictionary's definition: “*the state of being aware of and responsive to one’s surroundings.*” Does this respond to the question of what consciousness is? No.

In contrast, if you search “*algorithm*”, which is something that almost everybody in the information-dependent world we live in knows about or deals with today, you get about 365.000.000 results (roughly the same as consciousness). The first definition google gives you “*a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.*” Does this answer the question of what algorithm is? Yes.

Now, if you try searching “love” you get about 11.240.000.000 hits. And also in this case *nobody knows what it is.*

This introduction has of course no empirical and even less epistemic intention or value whatsoever, and only serves the purpose of dramatically enhancing the fact that although not being able to really answer satisfactorily to that seemingly simple question, a lot about it can be written. Consciousness has of course been studied thoroughly, however many questions still arise, making it some sort of *scientifically intractable problem* for now.

As Oizumi et al argue, “Understanding consciousness requires not only empirical studies of its neural correlates, but also a principled theoretical approach that can provide explanatory, inferential, and predictive power.” (Oizumi et al, 2013, p1).

Behaviorists generally used to focus only on what can be directly observed and measured, which of course refrained from analyzing consciousness in order to maintain objectiveness in their scientific enquiry. But the last couple of decades some key steps turned *the science of consciousness* into something that should and can be *objectively* studied. The ‘dominant’ principle of objectivation has shifted into a *principle of subjectivation*, in order to understand both the external and the internal world. (Hunt, 2020). Intersubjective confirmation is a new standard replacing “objectivity” meaning: all “objective data” are discussed and decided in a scientific community whether it is accurate and relevant before labeling as true. “Truth is intersubjective, not objective. There is no *view from nowhere*. There is always a somewhere, a perspective, a subject.” (Hunt, 2020). This constitutes what is known as the *epistemological hard problem*, which interrogates about the data and the questions that have to be made in order to unravel the nature of consciousness. It somehow stands behind the today famous *ontological hard-problem* announced by David Chalmers to reframe the classic mind/body problem (Chalmers, 1996), which has been criticized by other prominent authors such as Daniel Dennett, as will be discussed further on in this section.

The stakes are well captured by Edelman and Tononi: “Consciousness poses a special problem that is not encountered in other domains of science. ... What we are trying to do is not just to understand how the behavior or cognitive operations of another human being can be explained in terms of the working of his or her brain. ... We are not just trying to connect a description of something out there with a more sophisticated scientific description. Instead, we are trying to connect a description of something out there—the brain—with something in here—an experience, our own individual experience, that is occurring to us as conscious observers.” (Edelman and Tononi, 2001,p11). It seems that to unravel consciousness (or get closer to it) we have to resolve a dialectic going on between inside and outside reality, or more eloquently put between an *intersubjective* measurable world (objective), and a more hard to measure subjective internal world. And how do these two relate and where do physical structures bridge this dialectic?

In the late fifties Erwin Schrödinger wrote a book called *What is Life: The physical aspect of the living cell & Mind and Matter* where he approached this *hard problem*, recognizing that objectivation is just a step toward understanding nature but that science must progress and include subjectivity. “By [the principle of objectivation] I mean ... a certain simplification which we adopt in order to master the infinitely intricate problem of nature. Without being aware of it and without being rigorously systematic about it, we exclude

the Subject of Cognizance from the domain of nature that we endeavor to understand. We step with our own person back into the part of an onlooker who does not belong to the world, which by this very procedure becomes an objective world.” (Schrödinger, 1958). [§2](#)

Neuroscientist Christof Koch makes a reference to the fact that based on the best available information we have we tend to make inferences, *reasonable inferences* which he underlines to be a kind of tool broadly used in science (Koch, 2019). Neuroscientists and cognitive philosophers have recently been developing tools that depend upon *reasonable inference to measure consciousness*. The world is a creation of our own minds based on the *imperfect sense data* we process from the outside objective world. As such, even though our minds and our subjectivity are complex, they are also a part of nature, not being distinct from it in a qualitative sense, and so they do have to enter in the *consciousness equation*. [§2](#)

How can one actually know anyone or anything to be conscious, since the only consciousness one can be sure of is its own? *The only consciousness I can know with certainty is my own. Everything else is inference*. As stated above, tools for *measuring consciousness* are being designed. Some of these focus on integrated information as a measure of consciousness as argued by neuroscientists Giulio Tononi and others such as Christof Koch. Integrated information advocates claim that types of consciousness can be measured by amounts of bits. For instance a light diode containing only one bit of information to vary between two states (on & off) would count as a very limited type of consciousness. (Hunt, 2019). This might turn out to be an interesting way to measure degrees of consciousness levels and will be later on addressed again in the *Reasonable Inferences* section. [§3](#) Nevertheless, the fact that we cannot measure consciousness to actually determine the extent of which we ourselves are conscious, strongly undermines our ability to determine conscious states of others, whether they are living creatures or computerized *entities* of some sort.

3.1 _ Uploading Consciousness

For the time being it seems that we can *reasonably infer* that we acknowledge each other's consciousness states mainly by observing and comparing what we recognize as conscious behavioral patterns. As for our own consciousness, we simply feel conscious

in the sense that we are aware of our own subjective experiences and that make our own feeling, hearing, seeing etc *really our own*. “These conscious experiences lie at the heart of our mental lives, and are a central part of what gives our lives meaning and value. If we lost the capacity for consciousness, then in an important sense, we would no longer exist.” (Chalmers, 2010, p35). So in light of this reasoning, Chalmers argues that if an upload would not exhibit consciousness (if we were to measure this somehow), then *its existence* would be questionable and if it were to exist in this *consciousnessless* form, then it would definitely be a zombified kind of entity. Consciousness does give meaning to life, even if we do not have yet the means to measure it, and this too will be further discussed in the Thought Experiments Applicability section, for there are many authors (*or perhaps even all humankind*) that believe it to be *essential, regarding the meaning of life*. [§3](#)

So objectively one needs to pose one of the major questions regarding MU, which is to know whether or not the upload can be conscious. And how can we determine this?

Although neuroscience keeps unveiling consciousness correlates we still do not know how – and not less importantly why – consciousness even arises. Yet, and following Chalmers’ argument, we take *its existence* for granted. However, as he claims, “There is nothing even approaching an orthodox theory of why there is consciousness in the first place. Correspondingly, there is nothing even approaching an orthodox theory of what sorts of systems can be conscious and what systems cannot be.” (Chalmers, 2010, p35). We do know though, that it is the brain that gives rise to what we call consciousness. Or at least we *reasonably infer* so. But knowing how the brain functions and knowing its physical structure unfortunately does not suffice to explain conscious experiences, or a system’s subjective experience. A classical example is that you can know all about colors which still does not mean you can for instance experience *seeing* the color red. Or, to take another example, knowing all about mice does not mean knowing *what it is to be a mouse*.

Yet, the consciousness issue is of course a major concern in the sense that if an upload will not exhibit consciousness whatsoever, then it would downgrade MU’s interest to a great extent. Hence, and although still being an open question, it is of the utmost importance to investigate it, which is why a large community of neuroscientists and also many of the most prominent contemporary philosophers have become quite engaged to it in these last two-three decades; McGinn (2004), Searle (1980, 2004), Chalmers (1996,

2010) and Dennett (2012, 2017) to name just a few. *Biological theorists* of consciousness argue that consciousness cannot be non-biological whereas *functionalists* advocate that the causal structure and role, the organization of function, are the key in establishing consciousness, instead of the underlying biology. A non-biological system, or an upload, could be conscious in the functionalist perspective as long as holding the adequate organization pattern for it. Again, and underlining the above claims that we do not know if a non biological system such as a computer could in fact become conscious, we also do not know how a biological system can! Moreover we do not know what kind of differences between these systems there are that could trigger a conscious state, which leads some to assume that both systems should in principle elicit consciousness. (Chalmers, 2010; Bostrom, 2014; Kurzweil, 2005; Koene, 2013; Wiley, 2018; Hayworth, 2012)

An ideal upload would be isomorphic in the sense that every component it was made of would replicate in the exact same form its counterparts of the original system, always maintaining the processing of information, the whole cognitive or computing procedure just the way the original brain would have produced. This upload would then be called a functional isomorph (*of the original system*) as long as the components (for instance neurons) of the original system are in fact able of computing, which in principle they are, as discussed at the end of *the Analogical Inferences* of the previous chapter on WBE.

To better exemplify this concept, many authors (Chalmers, Häggström, Cerullo) recur to the gradual uploading process (such as a nanotransfer) because it seems to be more acceptable than the so called scan-and-copy uploading approach, that will be addressed in particular in the section 4.7 of this chapter (*Metaphysical Equivalence of Personal Identity in uploading scenarios*) since there is some evidence to argue in favor of an equal acceptance of both approaches. Nevertheless, and as stated for this scenario, all components, whatever they might be, are to be uploaded one by one over a period of time so that at the end of the procedure they will emulate with precision the component they replaced and interact in the exact same way the original system did. At the end of the replacement procedure, no original component is left and the whole brain would have been emulated *without even knowing* this procedure to have occurred. In the end we should then assume that if the original brain is said to be conscious, the *perfectly replaced* one should too (along with keeping memory of its prior identity)! So now the question arises as to what happened to what we call consciousness during the

transferring procedure. During the time that went between the first 1% of replacement until the last 99% of uploading, was that consciousness kind of slowly migrating? Did the transition occur fully right at once, as soon as merely one component would have been replaced? Did it gradually fade out of the original system? Thought of in terms of percentage, would the system with *more* percentage of the mind being uploaded somehow be the *more conscious* one? Or would it be conscious throughout the entire procedure?

Between gradual fading and sudden disappearance, as soon as just one element would have been replaced, the former seems more plausible just because there are even today records of small replacements or removal of small components that do not affect the subject's consciousness level. So the latter hypothesis of sudden disappearance seems quite improbable. But even the gradual fading case does not seem that probable at all, for there would have to be a kind of intermittency of consciousness, a kind of come and go of it as it would gradually fade out and if we compare this to clinical situations such as anaesthesia or split brain syndrome cases, this doesn't seem to happen. Chalmers, 2010). [§1](#) Empirically, either there are conscious states or there aren't any, but flickering in and out does seem a bit far-fetched. One minute you would be conscious and the next act like a zombie, and this repeatedly, so the most probable hypothesis would actually be that of not having consciousness at all or, on the contrary, retaining consciousness throughout the process. As Chalmers claims, "On this view, all partial uploads will still be fully conscious, as long as the new elements are functional duplicates of the elements they replace. By gradually moving through fuller uploads, we can infer that even a full upload will be conscious." (Chalmers, 2010, p39). David Chalmers argues that this will happen even when we start by a partial upload. Imagine just a few components (suppose 1%) of the original mind being uploaded. It would make sense that after some time of exhibiting a normal conscious state – one that people would recognize as being so, i.e., of being conscious – this would therefore also convince people that a total upload (100%) would be able to preserve consciousness. But he then leaves out the *logical sequelae* one could delve into. For instance, that functional isomorphs would also have qualitatively identical states of consciousness, which he calls the *dancing qualia* (Chalmers, 1996), a subject that will be further ahead addressed in more detail. [§2](#)

Nevertheless, and given the hypothesis that functional isomorphs would be able to preserve consciousness (*even if not the identical conscious states*), then one could label

consciousness as an organizational invariant, in the sense that systems exhibiting the same patterns of causal organization also exhibit the same states of consciousness independent of the substrate, whether biological or made of silicon for that matter, hence expected to be simulated in a computer. So, as Chalmers argues, “given that consciousness is an organizational invariant, we should expect a good enough computer simulation of a conscious system to be conscious, and to have the same sorts of conscious states as the original system.” (Chalmers, 2010, p40).

As stated earlier, we judge others to be conscious (or not) essentially on their behavior, which does not mean that they are (or are not) conscious because they could just be acting as if they were (which actually is held as *one of the flaws* of Turing's Test, and also the motive for Chalmers' thought experiment regarding zombies). Accordingly, if we judge people to be conscious or not by their behavior, then we should do so with uploads if they were to exhibit the traits we associate to conscious states (even if unproven). This is the so-called Principle of Assuming the Most: “Assume that any emulated system could have the same mental properties as the original system and treat it accordingly”. (Sandberg, 2014, p15), with which I concur.

Now, one way to argue for the probability of consciousness in MU is the so-called *Computational Theory of Mind* (CToM), which states that what matters for consciousness is not the material substance itself but its organization with the right kind of information processing. Various of the aforementioned philosophers (and others) refute this claim by putting forward arguments such as “it [the CtoM] works only by denying the existence of what you and I and everyone else understand consciousness to be (that is, subjective feelings)” (Fukuyama, 2003, p167); and “consciousness is a biological phenomenon” (Pigliucci, 2014, p119); and the famous chinese room thought experiment of Searle (1980) which is a *reductio ad absurdum* kind of argument.

Searle assumes the CToM to be right in order to prove it wrong. We are supposed to imagine a room full of papers teaching Chinese, with Searle himself in the room. The papers teaching Chinese correspond to the program that gives the instructions to the computer to speak Chinese, in the analogy. There are also two windows in the room, which Searle can use to give and take information very much like input/output in case of the computer. Searle would play the CPU role in this analogy and, according to CToM, he should then with these instructions be able to speak Chinese, which he labels as absurd since one thing does not directly mean the other and he certainly cannot speak

Chinese. Daniel Dennett (1981) and other philosophers replied to this thought experiment arguing that it is not Searle that learns Chinese, rather the whole room which consists of the CPU (Searle) plus the instructions (papers). So the system is responsible for learning Chinese and not Searle himself and a series of debates began between Searle, Dennett and others (which I cannot analyze in length here, as they are beyond the scope of this thesis). This discussion proves the controversy generated among thinkers regarding consciousness and the possibility of CToM that, either way, has not until now really been proven to be true, but also not proven to be false which gives it some room in terms of possibly explaining how consciousness arises in the physical world. (Häggström, 2016).

3.2 _ Debating *the Hard Problem*

The *hard problem* consists of investigating how awareness can actually arise from matter. Consciousness is a context-dependent concept. Self-awareness, being aware of one's existence, knowing that there is a self or experiencing sentience, are all definitions of what David Chalmers calls *the easy problem* as opposed to consciousness' *hard problem*. They are not necessarily easy to explain, in fact they may be very difficult, but the fact that they *can (or could)* in principle be explained, through *conventional and empirical science*, makes them an easy problem whereas explaining how the brain generates consciousness altogether is really where the difficulty lies. Science is objective, and *studies and investigates* things objectively, meaning that the subjective experience is to be left out, resulting in a *problem* when addressing consciousness. Schrödinger was aware of this, as duly noted earlier in this section. Chalmers, as also other scientists and philosophers, feel that there is an *explanatory gap* between the neurobiological processes and *experiencing them*. There is some kind of a bridge between *traditional science* of the objective physical description of those neurologic processes with the subjective description of the experience *per se*. (Chalmers, 1995, 1996 & 2010).

That bridge could lie in *information*, as will be addressed further ahead in the *Reasonable Inferences* section of this chapter. [§2](#) & [§3](#)

On the other hand, Daniel Dennett calls Chalmers' so-called hard problem of consciousness the easy problem of consciousness since he considers it to be a bundle of tricks that consciousness plays on us that are perfectly explainable without recurring to any revolutions in physics or emergent novelties of any kind. He concedes not being able to prove that there really isn't a hard problem as described by Chalmers but also conversely, there is no way of proving there is one. He considers that once we unravel these tricks as some kind of *user-illusions* instead of a bunch of mysterious phenomena, we then start understanding how the brain creates consciousness: "(...) consciousness is a physical, biological phenomenon like metabolism or reproduction or self-repair that is exquisitely ingenious in its operation, but not miraculous or even, in the end, mysterious." (Dennett, 2003, p7). He also considers that there is a *spoilsport* of some sort in explaining consciousness, which relies in the fact that people tend to make it more mysterious than it actually is. So if one would be able to explain it, then it would somehow diminish its mystery, and also its importance.

Dennett goes through a lot of examples stressing the power of suggestion the brain has, like the *déjà vu* effect and *belief of expectation*. Events we sometimes think to experience a second time when all that eventually happened was an interruption of the perceptual process, splitting it into the past just a few milliseconds, and making us believe to have experienced it before. *Déjà vu* could only be a delay in processing rather than an esoteric hyperdimensional experience we tend to make of it as Dennett claims: "...a familiarity detector (or, alternatively, a novelty detector) that discriminates all incoming signals into those that are novel and those that have been encountered before. (There is evidence that the hippocampus has this very task among its duties, so this is not an entirely gratuitous speculation)" (D.Dennett, 2003, p10). Yet from the inside, from our own first person perspective, the signals (and the delay between them) would be indistinguishable making us believe to have experienced it a long time ago (or even in another life or anything more esoteric) whereas in reality it just happened a few milliseconds ago. *Belief of Expectation* is another kind of power suggestion the brain can make on our consciousness, everytime the brain fills in the brackets (so to speak), when it finishes the job that our own visual experience started: "When the brain takes the suggestion, the brain is forming a belief or expectation, not painting a picture for itself to look at." (Dennett, 2003, p15). As in the kind of experience of finishing wrong written sentences, the brain delivers exactly what is expected, because it already has some prior reference for it: "Magic reveals how wrongly we remember what we have seen, discloses the way

in which memory is the bearing and nursing mother of illusion. Memory is the magician's assistant, confederate, and shill.” (Dennett, 2003. P15). This is what Dennett calls heterophenomenology. (Dennett, 1991): “People undoubtedly do believe that they have mental images, pains, perceptual experiences, and all the rest, and these facts about what people believe, and report when they express their beliefs are phenomena any scientific theory of the mind must account for.” (Dennett, 1991, p98).

Ultimately, one possibility is to drop the subject or the self when explaining consciousness: „All the work done by the imagined homunculus in the Cartesian Theater must be distributed around to various lesser agencies in the brain. Whenever that step is taken, however, the Subject vanishes, replaced by mindless bits of machinery unconsciously executing their tasks. Can this be the right direction for a theory of consciousness to take?” (Dennett, 2003, p15).

As stated earlier, Chalmers argues there is a hard problem and that theories explaining the functional interdependencies (the backstage machinery of the brain so to speak) can only solve *the easy problems of consciousness*. (Chalmers, 1996). Dennett argues that this hard problem is a *tempting idea*, but also a mistake, although he cannot prove it to be wrong: “Once we appreciate all the non-mysterious ways in which the brain can create benign user-illusions, we can begin to imagine how the brain creates consciousness.” (Dennett, 2003, p19).

4 _ Personal Identity

Whether or not there is a hard problem of consciousness, or even if an upload should become conscious, there would still be the personal identity issue, that may turn out to be an even *harder problem* than consciousness itself. Many authors use teletransportation as an analogy to MU when approaching the personal identity issue. Although not being the same thing as a mind upload – teletransportation aims to preserve physical organization whereas uploading intends to preserve functional organization and in another kind of substrate – the issues they raise in terms of identity transfer are closely related. This has been widely explored in science fiction for many years and also various philosophers have entertained a couple of thought experiments in the subject (Lem, 1957; Parfit, 1984, Olson, 2010). Many of the current approaches

when discussing uploading show a big influence of Parfit's *Reasons and Persons* published in 1984. Although he does not discuss uploading in the book (given that he was writing in the early 1980s, when the issue was still not being explored), he does review the phenomena related to teletransportation, whose possible outcome scenarios can to some extent also be applied to MU. Derek Parfit thoroughly explored identity in that book using thought experiments for a number of scenarios ranging from split-syndrome brain, swapped memory and personality to teletransportation, all of them at a time very much influenced by science-fiction and philosophy. At that time MU was just an idea, at most a vision, but not yet a subject. A couple of decades later technological exponential growth caught up science-fiction but his classification remains a kind of guideline when approaching the issue. According to Parfit, there are two types of personal identity (Parfit, 1984):

- “qualitative identity” which is *the measure of similarity* between two minds (and here we should note that Parfit’s term is not self-evident).
- “numerical identity” which should answer the question: will my upload be me (*or just a copy of myself*)?

Numerical identity thus implies qualitative identity to somehow be incomplete inasmuch as for it to be “real” it needs a consciousness continuity of some sort, or something that could *move from one mind to another without* loss of information. None of these are measurable, for now. However, AGI eventually developing (or exploding) into a superintelligence of some kind may deliver the tools for measuring it or simply clear things up by showing it to be an illusion. This will be further addressed in the forthcoming chapter (Superintelligence - Part IV).

There is an intimate relation between survival and preservation of personal identity because one could survive teletransport (if it were to be possible), as uploading for that matter, and yet, not preserve personal identity. Asking the subject after going through such a procedure would not help very much because she would always claim to be the same person for it does not have any reference to a supposedly previous one, hence every teletransported subject (or uploaded person), would claim to be the same person (Häggström, 2016; Chalmers, 2010; Chalmers, 2020, etc).

When approaching uploading scenarios it is not infrequent to confuse survival and identity preservation. However it is quite unambiguous that also our identities change over our lifelines. Thus, for instance, a child's identity is not exactly the same as that same person as middle aged adult and surely not when in old age. So there isn't really a mathematical identity even in ourselves, at least not over some greater periods of time. But when addressing survival outside the uploading scenarios, in terms of identity there is no room for discussion anymore because *children don't die to transfer their identity* on to their middle-aged counterparts and further on. Also, contrary to survival, identity is not addressed in an all-or-nothing way. We could consider degrees of partial identity insofar as the middle-aged subject does relate more with the child than the old-aged one does. But the criterion of survival has been put forward in all these discussions and is usually applied in a binary fashion in the philosophy of mind literature, such as Parfit's and Chalmers'. As Wiley & Koene maintain, "identity might be judged to partially preserve via a destructive mind uploading procedure while at the same time the person is judged to either survive or die." (Wiley & Koene, 2015, p4).

Elizer Yudkowski, a computer scientist at MIRI (Machine Intelligence Research Institute) made a thought experiment in 2008 regarding non-destructive teleportation, on purpose, as opposed to the destructive hypothesis only to underline his view on the subject, which then later on triggered some thinkers to argue for and against it. Yudkowski's argument consists of holding that in this case, and precisely because it supposedly was to be non-destructive, *the original* is taken to stay alive and therefore relegates the *teleported subject* to the status of a copy. He then states that in the destructive teleport hypothesis, one would have to consider murder and birth at the same time (the original would have died and the copy would have been born), which *obviously* forces the teleported subject *to really be the copy*. (Häggström, 2016). This was precisely the view taken by Pigliucci (one of the proponents of the disclaimers of MU, that will be addressed below in the critics and criticism section of this chapter) claiming that "if it is possible to do the transporting or uploading in a non-destructive manner, *obviously* we are talking about duplication, not preservation of identity" (*italics in the original*). (Pigliucci, 2014, p128). Pigliucci highlights the fact that this would constitute a duplication, hence not a preservation of identity because the original subject would have survived in the non-destructive hypothesis. This reflects the intimate relation between survival and personal identity, and if we consider that preservation of identity is *all about* survival then we can do some *reasonable inferences*.

To start, it seems quite ludicrous to assume that this issue would depend on the atomic precision insofar as I acknowledge that the very own substance I am made of, atoms, molecules, electrons, particles generally speaking, keep changing on a regular basis. So I cannot expect an upload of mine to be made of exactly the same atoms in order to accept that MU would have preserved my personal identity. Therefore it is reasonable to discard ‘property’ as the sole responsible for not preserving identity in an upload (or teleport). Otherwise I would have to admit to being a new person every now and then, depending on the turn-over of cells, molecules etc... It seems another reasonable inference to make that although changing the molecular set over time, we do preserve a pattern of organization very much like that water makes in a stream as it flows. It slowly changes the elements it is made of, and in a *continuum*, but always preserving a pattern over longer periods of time. (Häggström, 2016).

The functional isomorph hypothesis postulated earlier in the previous section, would tell us that consciousness could have been preserved (*therefore survive*) in the upload (and in the teleport too, for that matter), for even in the case of the copy it would entail the same organizational pattern as in the original subject. Yet, since personal identity cannot be an organizational invariant, simply because the copy and the original, even being qualitatively identical cannot be numerically identical - it follows that one cannot infer MU to preserve identity *simply* by preserving the pattern. (Chalmers, 2010).

If we are to assume that in a non-destructive case there is duplication, and the upload does not preserve identity, then we would also have to assume the same in the destructive scenario, and both subjects could not share the same identity. This is the most common view of those authors who are more skeptical of MU.

Following the kind of rationale the majority of these authors use when approaching the different possible uploading scenarios, I will also display them in the first person, as if I would undergo such a procedure in the near future:

- a) If I would submit myself to such a procedure, I would first want to know how my survival chances stand. So in the non-destructive approach I would definitely survive only not knowing there would be two of me, or me and a copy of myself, whatever that would mean. The fact that I have already written in this option “me and a copy of myself” basically puts me together with probably the majority of people (if not really everybody else), that are rightfully convinced they can

acknowledge themselves as biological entities in possession of what they (and I) believe to be their (and mine) personal identity.

- b) If I would submit myself to the destructive approach, my biological me would *by default* not survive, and I would have to just accept what I now still regard as *only a copy of myself*, for I am not sure if it would have what I consider to be my personal identity, even if I would be reasonably convinced that it could in fact be a conscious copy of myself. I would however undergo such a procedure, if for instance I would find myself with a terminal disease or any other reason that by then could convince me of at least having a *continuity of myself* of some sort.

So the main question here is always that of personal identity preservation. This is what I consider to be *my survival*. As already discussed, an atomic precise identity to guarantee preservation does seem far-fetched. I would definitely not think that for my personhood to persist over time it had to be with the exact same atoms and molecules, for I'm aware that even today, as biological as I am, they keep changing and I am reasonably convinced to still remain *the same*. On the other, and although preserving consciousness might eventually be feasible, that does not comfort *my survival* pretensions entirely for it still would not directly mean preserving my identity. Nobody (*or almost nobody*) would be willing to risk the chance of undergoing destructive uploading without guaranteeing identity preservation if they would tie it to survival, even if the *continuation of consciousness* could be achieved. "Most theorists hold that a certain sort of *continuity* or *connectedness* over time is required for survival. But they differ on what sort of continuity or connectedness is required." (Chalmers, 2010, p41).

4.1 _ Main Personal Identity Theories

This sort of connectedness, and its continuity, could very well be linked to the space and the time mediating between the original system and the emulated one, which I will call the *space-time continuum** between both of them. On this kind of connection there are currently mainly three sorts of views or theories. The *biological* view states that personhood survival requires the biological brain or organism whereas the *psychological* view claims that survival of identity demands a psychological continuity over time preserving memories and causal related mental states among others. The third view /

theory of identity, holds that preservation of identity can be achieved through the most closely related subsequent entity so that consciousness may continue the most identical to the original. This so-called *closest-continuer* theory is probably the *favourite* of MU proponents generally speaking. “Closest-continuer theorists are likely to hold that the answer depends on whether the uploading is destructive, in which case the upload will be the closest continuer, or nondestructive (in which case the biological system will be the closest continuer.” (Chalmers, 2010, p42). [§2](#)

*No author, other than me, uses or has used this term in reference to these MU issues. It is only a simplification for reasons of clarity. Space (the distance between the original and the upload) and Time (the period between the original and the upload) do seem to cause reluctance in proponents against MU. It is duly addressed below in the critics and criticism section. It can be confused with the so-called space-time continuum, and that does not seem to be a problem because there may even be a connection to the kind of physics, particularly quantum physics, when considering consciousness.

If one of the main phenomenal or maybe metaphysical issues separating the two hypotheses, non-destructive and destructive (whether teleport or upload), is a kind of interruption in the earlier mentioned *space-time continuum* (a discontinuity in the connectedness) then it definitely empowers the survival issue in the personal identity equation. Because it does seem acceptable that if both subjects would survive in the non-destructive scenario and the procedure leading the teleported one (*or uploaded*) to be *merely* a copy, then we should also take this for granted in the destructive scenario. If I would not *survive as a copy* in the non-destructive scenario why should I *survive as a copy* in the destructive one? (Chalmers, 2010). [§1](#) Furthermore it also highlights the *space-time continuum* to be the challenge to overcome in order to allow a MU to succeed. Again, this particular space and time issue, or discontinuity of connectedness issue, will be duly discussed below in section 4.7 of this chapter (*Metaphysical Equivalence of Personal Identity in uploading scenarios*).

4.2 _ The neo Ship of Theseus

If in the gradual uploading scenario I would replace small components of my brain with functionally isomorphic nanochips for a certain period, starting with only 1% until the

whole system would have been fully 100% substituted (as was already considered in the consciousness section only now regarding personal identity) *would I be myself* throughout the procedure? Would I cease being me after just 1%? Or would there be some kind of threshold figure (for instance 51%, just for sake of argument) after which *I would end and another me would start?*

It definitely seems awkward to reason that replacing 1% of my brain components, for instance a bundle of neurons of whose functionality and its whole connectome were perfectly known and reverse engineered so to speak, that once emulated, would have such a drastic effect *and I would simply stop being me*. If this should be true one could also argue that we gradually stop being ourselves during the course of our normal lives. “...insofar as we ordinarily survive over a period of many years, we could survive gradual uploading too.” (Chalmers, 2010, p45). Chalmers strongly argues in favor of the gradual uploading approach and presupposes that the system should be active throughout the process ensuring continuous consciousness: “Then we can argue: (i) consciousness will be continuous from moment to moment (replacing a single neuron or a small group will not disrupt continuity of consciousness), (ii) if consciousness is continuous from moment to moment, it will be continuous throughout the process, (iii) if consciousness is continuous throughout the process, there will be a single stream of consciousness throughout, (iv) if there is a single stream of consciousness throughout, then the original person survives throughout. One could perhaps deny one of the premises, but denying any of them is uncomfortable. My own view is that continuity of consciousness (especially when accompanied by other forms of psychological continuity) is an extremely strong basis for asserting continuation of a person.” (Chalmers, 2010, p45).

4.3 _ Branching Identity

Yet none of the three main current identity theories answer satisfactorily to the uploading scenarios envisioned. If or when possible, MU will require new identity theories as was stressed out by many of the experts interviewed for this thesis (see [SSI-Critical Analysis](#) - Part V, [Question IV + Analysis #8](#)). But until then other theories are beginning to develop like the Branching Identity Theory advocated by Michael Cerullo. If we define identity as continuity of consciousness or of subjective experience within the same entity

through time, then personal identity could branch into multiple copies where each and every one of them would allow continuity of consciousness with the original. But this forces us to dismiss the usual stance that many thinkers advocate, namely numeric identity, which by default only allows one *numerical identical thing* to exist. Cerullo admits this to be unintuitive but argues that this alone cannot be used to exclude this theory only because it does not fit well in our *primate brain*. He argues that there are lots of other unintuitive concepts (such as for instance the space-time dimensional concept) we had to acknowledge throughout history and humankind's evolution. Furthermore he argues that we seem to easily imagine past-scenarios where things could have turned out differently had we gone alternative routes. So in those cases we imagine different selves from our present self quite easily. I would complement this reasoning of Michael Cerullo with a rationale given by Robin Hanson when interviewed for this thesis, and quoting from the transcription of the interview with him (see Appendix L, Robin Hanson, 2020):

“When we look backwards, and we see the path that leads to us we see that we are willing to embrace all of that like us. When we look forward to things that we might become in the future we are less clearly willing to embrace that as us. Post-humans as you might call them will just call themselves humans. And they will look back and see them as a natural extension of their heritage just like you and I call even if our ancestors might not.”

“It's wrong to state that the future is more dichotomous than the present or the past. It's a natural intuition people have based on something called construal level theory. Construal level theory is a natural way in which we switch between abstract distant things and up-close concrete things, and our habit when we go into an abstract thing is to think that it's more dichotomous, that categories are more distinct, more internally consistent and more distinct across different categories. That's just a more natural way to think when we think of abstract things and when we think about the future, we tend to think very abstractly...”

So Cerullo claims that thinking about alternative history is very much like thinking about branching identity hence we have no hard time in accepting that an alternate self shared continuity of consciousness with our current self the moment before they began to diverge. However, and to complicate Cerullo's argument a bit, this is not exactly the

same as imagining other *contemporary selves beyond ourselves*, which of course naturally becomes unintuitive to think about consciousness splitting off into multiple equal selves. As he duly admits, this would be somewhat like imagining that two things are one. Nevertheless, and to his favour, consciousness, as abstract as it is, eventually does not behave as regular physics does and dazzles us when approaching the continuation of subjective identity. Cerullo provides a formal definition to the *instantaneous psychological branching theory*:

“There will be continuity of consciousness between any conscious entity P1 at time T1 and P2 at time T1 if P2 contains half or more of the psychological structure of P1 at time T1 and is activated at time T1.” (Cerullo, 2014, p29).

Branching Identity overcomes many of the obstacles encountered with the other three theories when approaching uploading scenarios and also to present-time issues such as the split-brain syndrome. It predicts continuity of consciousness in non-destructive instantaneous uploading allowing two identical entities to emerge and both of them (original brain and upload) sharing continuity of consciousness. (Cerullo, 2014). He believes this theory to present the best possible solution to the traditional problems identified by philosophers regarding personal identity in all uploading scenarios and identifies branching of consciousness as the major hurdle to overcome, but even in this case he claims there to be enough empirical evidence holding strong to the split-brain syndrome cases. [§1](#) “If we accept functionalism and the best current neuroscience (e.g. no special quantum process or non-computable processes necessary for consciousness), then branching identity tells us that uploading is an acceptable way to allow for the continuation of consciousness.” (Cerullo, 2014, p33).

In scenarios where branching could not by principle be applied then it would automatically reduce to psychological identity, thus still predicting preservation of identity. Such an example is the case of Chalmers’ gradual destructive uploading. According to Chalmers in this scenario the person would preserve consciousness continuity throughout the process but gradual destructive uploading does not allow for branching to occur, hence psychological identity would take over, as predicted. (Cerullo, 2014). However, Cerullo criticizes that gradual destructive uploading becomes the dominant scenario for MU, since it would require a foreseen advanced kind of molecular nanotechnology clearly not available yet. He too, as others already mentioned (Koene, Wiley, Hayworth), feels that scan-and-copy, together with brain preservation, is the more

likely short-term scenario to become available. *Branching Identity* removes any difference between gradual and instantaneous destructive uploading.

4.4 _ Iterating Human Predictor Response _ (CaseScenario 1 to 3) - I

My thought experiment *Iterating Homo Sapiens Intelligence* (MU+SI Thought Experiments, Part I, 3b) was presented to the experts interviewed for this thesis (see [SSI-Critical Analysis, Question IV + Analysis #7](#), Part V). Nevertheless, *Iterating Human Predictor Response* (MU+SI Thought Experiments, Part I, 3a) shares a common background with it and will be here addressed relative to Personal Identity, whereas the other one will duly be analyzed in the next chapter of this thesis - Superintelligence, Part IV.

In respect to these TEs, and on the negative side, most of the interviewees (Koene, Sandberg, Agar, Barrat, Cattell, Pierce) mentioned their ethical vulnerability inasmuch as dealing with *possible conscious uploads*, subjecting them to the experimental simulations in a variety of environments, pretty much feels like experimenting on real people, as long as one believes them to have actually successfully generated consciousness. If this were to be the case, then MU would have proven to be a successful technology and, of course one could not arbitrarily subject them to any kind of scenario. However, and depending on the scenario, still, a million fold of possible harmless scenarios might be worthwhile considering. With AI's help, building the right *evolutionary algorithms*, safe environments could be applied and a lot of frequently occurring errors, a lot of mildly adverse situations we're usually subjected to and a lot of unpredictable hazardous events could be explored and hence in the immediate future avoidable. So, on the positive side, such an evolutionary kind of approach could be very useful, as the interviewees also acknowledged. For instance Barrat: "But that [in reference to the TE] seems more plausible than raising an AI in isolation and trying to program it with ethics. It seems like if you could raise it in an environment where ethics, where kindness and other virtues were rewarded then you would have this hard wired nice thing instead of letting all the commercial drivers evolve it." (Barrat, [Question IV + Analysis #8](#), Part V - SSI, 2020).

Moreover, some of the interviewees made a clear reference to these TEs potential use when addressing opaque systems, as a brain emulation might definitely be the case (Sandberg, Loosemore), and MU, as these TEs argue, could bypass that black box and

unravel what is traditionally hard for us when addressing these systems. Furthermore, and as mentioned above, choosing the adequate kind of unharmed scenarios, and subjecting them to our *consciousness-freed-uploads*, could also count as an *evolutionary* and safe kind of algorithmic approach, as the interviewees also highlighted. See for instance Hayworth: “How much of our common sense and general intelligence is tied to our consciousness? Can we have the type of really subtle common sense intelligence without bringing in things like suffering and the sense of self and things like that? And if I had to guess, I suspect that those are decoupled, so iterating under that scenario where you would decouple your general intelligence” (Hayworth, [Question IV + Analysis #7](#), Part V - SSI, 2020). And Loosemore: “If I understand which part of the brain is doing the actual thinking, the core thinking part of it, we could just take that part out and it has no longer the rest, so there is no longer a consciousness. So this is a unit test. This is a *subsystem test* and you don’t have the danger of building a consciousness.” (Loosemore, [Question IV + Analysis #7](#), Part V - SSI, 2020).

Now, specifically, the TE 3a argues that our predictor response can be iterated through a wide range of scenarios, thus augmenting our predictive capacity altogether. It is noted in that TE that this ability has in the past of humankind certainly accounted for an adaptive edge. Forecasting does give us the chance to better evaluate the near-term future and act accordingly. The TE explores the fact that today, as biological finite beings, we only have access to a limited number (very limited) of possible situations, conditions or scenarios that we could submit ourselves to, in order to better answer them next time around. The fact that we could submit our uploads to an astronomical number of possible scenarios widens our range of simulations, actually emulations for that matter. Now, besides incrementing the scenarios we could be subjected to, also the fact that we could have millions of copies running those scenarios at the same time, could provide us with a fine-tuned forecast of an also astronomical number of actions we would take, without actually going physically through them. These two variables (number of scenarios X number of copies) combined together at the same time, would fine-tune the cutoff in possible responses and augment our predictive capacity.

In practical terms, human behavior is rather stereotypical. Although of course there are unique and individual actions, one’s overall behavior can be seen as a sum of lots of previous failed attempts. Like in machine learning (also a neural net), the human mind learns to predict the next moments of observed experiences (Karpathy, 2015). Based

upon our past behavior we can extrapolate our future behavior. There are certainly common facts in the general population that have a bigger predictive value as opposed to those more unique facts that very much define our personal individual traits. Recording all possible data relative to one's behavior may help to increment the accuracy in predicting behavior in the future and with all kinds of available technological devices today that threshold could in the near future with AI's development easily reach an astonishing human behavior predictability. (Turchin, 2018).

Our human behavior is becoming increasingly predictable not only because traditional *data* like genetic and biological records are an ever more accurate source of information about us, but also because of the amount of digital footprint about us available. Internet usage is a distinctive part of our behavior currently spending a lot of time online, and large amounts of this activity is recorded. (Kim, Biessmann, & Lee, 2014). Deep learning based on EEG data was used to reconstruct visual images (Shen, Horikawa, Majima, & Kamitani, 2017). EEG was used to reconstruct words from a brain and it is expected that in 5 years it could be used to record thoughts on a smartphone (Jonston, 2017). Personal assistants are evolving into the so-called exocortex that will ultimately end up in a digital copy of the owner. Computers have already replaced our memory in every aspect of our lives. The list could go on and in fact it is becoming as vast as the internet itself, hence future AI will have a ridiculous amount of data on all *sources of individuality* available to be able to reconstruct this individuality along with the aid of other technologies such as MU. With an eventual AGI, or a kind of *Superintelligence*, we will definitely maximize our human behavior predictability. (Yampolsky, 2014; Turchin, 2018).

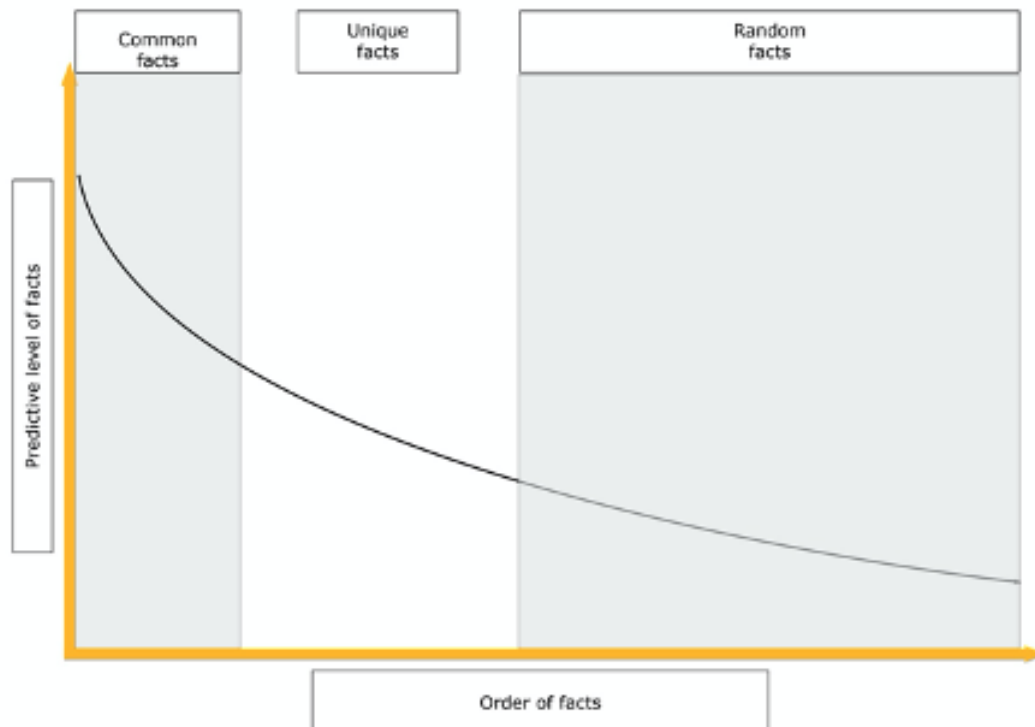


Figure 21 – Predictive level of facts. Unique facts form a rather narrow strip between common and random facts; inside this strip, the curve declines almost linearly and doesn't change much. (Source: Alexey Turchin, *Theory and Protocol for Indirect Mind Uploading*, 2018).

These random and unique facts (shown on Figure 21, taken from Turchin), could – as envisaged in the *TE Iterating Human Predictor Response* – be shortened and the curve more tuned with the common facts, thus augmenting our predictability of human behavior. Now this could be viewed as just another attempt of *human reductionism*, and therefore easily labeled as *inhumane*, or some other form of diminishing human experience. *Why would we ever want to preview our actions if they'd prevent us from experiencing them?* Well, think of all the planes that wouldn't crash, all the guns that wouldn't fire or just simply all those bad decisions that sometimes are made but that really could have been avoided. The balance is quite positive when comparing to maybe spoiling a romantic dinner or on the other hand, we would probably adapt and just maybe would learn to *make more of it*, whenever experiencing those elected and safe kinds of actions. By then *evolutionary algorithms* would eventually already be fine-tuned to the

point of filtering the kind of decisions we could improve without spoiling the experiences we would want to keep.

As envisaged in the TE, MU could enable us to preview probable actions based on our uploads running in a number of scenarios. Now, one could argue that it would be possible to do this even without MU, based on records and with the adequate AI. However those would be simulations, and not emulations, which makes a big difference. Furthermore, they would literally be avatars, but not actually *selves*, or even not exact duplicates of you, to keep Parfit's identity ideas as the reference. Again, also the ethical aspect could as mentioned above be safeguarded, since depending on the scenario, the upload might not have the conscious mode active as Loosemore also underlined in his rationale in his interview for this thesis (see [Question IV + Analysis #7](#), Part V - SSI). Or, conversely, if the conscious mode should be active for some particular reason, then the scenario would have to be tuned for it. A compromise between these two variables would easily be feasible technically speaking and most certainly ethically desirable.

It seems acceptable that there is nothing stopping us from researching and investigating pain systems without phenomenal consciousness. Analgesic practices used to reduce suffering can have counterparts in emulations. By default, WBE's technology should permit emulation's parameters to be finetuned to produce effects similar to those that drugs have in our real nervous systems, as Sandberg claims: "pain systems can in theory be perfectly controlled in emulations (for example by inhibiting their output), producing "perfect painkillers". However, this is all based on the assumption that we understand what is involved in the experience of pain: „if there are undiscovered systems of suffering careless research can produce undetected distress. It is also possible to run only part of an emulation, for example leaving out or blocking nociceptors, the spinal or central pain systems, or systems related to consciousness.“ (Sandberg, 2014, p17). Such an approach is in the digital case also possible to apply with more precision and even with reversibility, as opposed to biological systems.

4.5 _ Iterating Human Predictor Response _ (CaseScenario 4) - II

The last scenario of this TE argues about the possibility of *perfectly* matching two different persons' cognitive blueprints. Its occurrence is definitely of a low probability, however, even if only mathematically, of a non-zero probability. Hence, such a possibility,

however absurd it could be, would nevertheless have to shake current personal identity theories. As discussed in the [SSI-Critical Analysis](#)- Part V of this thesis, the interviewed experts believe (*and this seems to mirror the general intuition that most people will have when faced with this possibility, I might speculate*), this would mean very little since even if they were to be identical it would only be at that particular moment of conception. From there on, time would make those two persons have a different experience so they would end up diverging anyway.

However, quoting Sandberg “This shows that personhood and personal identity isn’t a real thing.” (Sandberg, [Question IV + Analysis #8](#), Part V - SSI). Also, Agar (among others) stressed in the interview the need for developing new concepts of identity. Loosemore highlighted an important detail that makes all the difference: “How much of a personality is the underline design and how much is the experience and knowledge?” (Loosemore, [Question IV + Analysis #8](#), Part V - SSI). And how sure can I be that experience would in fact make them diverge? I grant it to be very low probability, and even then, *experience acquired through time* would almost certainly have to make those two persons diverge. Again, quoting Loosemore on this particular issue: “I think the two of them would think in much the same way about any problem but that would only be the case as long as their past experience was not part of the solving problem.” (Loosemore, [Question IV + Analysis #8](#), Part V, *my italics*).

Yet, the non-zero probability remains, and to me, *contra-mundo*, it means it still is an open question. Because, and unpacking this last thought of Richard Loosemore, how certain can I be that *their past experience was not part of the solving problem*? Could it not be that those *predictive levels of facts* mentioned above in the first three scenarios of the TE, would in this particular case exhibit very much identical curves, if not even equal? So, in the extreme and very low-probability, would their unique, and maybe even random facts not make them experience life from there on in pretty much the same way? How sure can we be of it, without actually undergoing the empirical experiment of, say, following them throughout their lives, admitting such a case should in fact take place?

MU will definitely require a whole new revision on Personal Identity, and the final scenario of this thought experiment (CaseScenario4) shows just that.

4.6 _ Social Construction of Identity

MU has essentially been linked to the transfer of what we consider to be *the mind*. And most of the proponents of a successful uploading generally assume the mind to be the pattern of information that constitutes our brain's functionality, memory and our personality traits, basically that entails what we also consider to be our personal identity. (Moravec, 1988; Kurzweil, 2005; Hayworth, 2010; Goerzel, 2012; Rothblatt, 2012; Koene, 2011; Wiley, 2014). All of these proponents assume personal identity to be essentially driven by functionalist, psychological states and biological processes but there are also authors arguing the relevance of identities as being constructed socially, and then claiming this to be enough to sustain a collective belief in the persistence of identity. (Bamford & Danhaer, 2017). "What if personal identity is constituted in, and sustained through, our relations with others, such that were we to erase our relations with our significant others we would also erase the conditions of our self-intelligibility?" (Kompridis, 2009, p. 27). "...one could follow the lead of the social epistemologists and other externalists in epistemology and... argue that external factors are really the ones that fix the identity... [One could... try to locate the source instead outside of the subject's cognitive domain." (Hongladarom, 2011, p. 541). Bamford et al concede that these ideas are not that well accepted: "they go against commonly held notions that questions of continuity of personal identity are too important to be determined arbitrarily or to depend on third parties (Parfit, 1984, p. 267)." (Bamford et al, 2017). However, they argue that there is sufficient ground to accept that personal identity is also a social construction and that this should be taken into account when approaching MU scenarios as well. Contrary to the first-person's perspective, which of course is tied to the continuity of one own's subjective conscious experience, the third-person's perspective is different inasmuch as it relates that self's continuation as a bundle of social issues because *we all stand in relations to one another*. If an upload were to succeed, all these issues were to be rightfully accounted for in an identity transfer. Their social-constructed personal identity claim, essentially states that independently of the perspective (whether first or third-person), social events and arrangements affect that identity. "From a first-person perspective, his identity (the continuing subject that he experiences himself to be) is causally socially constructed." (Bamford et al, 2017, p13).

If certain properties of identity were actually to be constitutively intersubjectively constructed, then it does not seem absurd that besides memories transferred from the

biological entity these would definitely also be complemented or completed with memories taken from a collective belief over that specific individual. There are a number of qualifications in all kinds of relationships we have with other peers that characterize our identity and that ultimately depend on that collective belief. The socially constructed nature of our first-personal identities is a function of the embodied and distributed nature of cognition. (Kirsh, 2006; Clark, 2008; Menary, 2007). Our memories are filled with concepts we keep building continually over life making up our subjective history and that to some extent end up labeling our *self*. But we do this to others too, by interacting with them. “By extension, a personal identity consists in the set of concepts that relate to a particular person; importantly, this set of concepts resides not in one place but in the memories of both the human that the person is a conceptualization of, and of all their peers. This set of humans each have their own memories which relate to the person, and this entire body of information forms part of the personal identity.” (Bamford et al, 2017, p15).

Moreover, these memories and other cognitive capacities our personal identities are made of, also rely on social constructions made of cognitive artefacts in networks that include other people, our peers, which then also help to sustain and remember those memories. (Norman, 1991; Heersmink, 2013). The environment, all our surroundings, is filled with data *concerning us* and so can also act as a substrate for our personal identity. “Thus we can think of a personal identity as a continually evolving set of information, which is stored disjointly across a set of different substrates, typically with the human body in question as the hub.” (Bamford et al, 2017, p16). Furthermore, and although it is true that we automatically link our bodies to our personal identity it isn’t less true, that these associations are also made of decisions with the help of our social peers. [§1](#)

4.7 _ Metaphysical Equivalence of Personal Identity in uploading scenarios (or Scan and Copy Vs Gradual in-place Replacement)

As stated earlier, some preference has been shown for the gradual uploading scenario over the scan-and-copy one by many authors, essentially those exercising critique over uploading generally speaking (Hauskeller, 2012; Pigluici, 2012; Corabi & Schneider, 2012). Likely because it seems more intuitive to accept such a procedure than because

of technical reasons or even metaphysical, as other authors (Wiley, 2014; Koene, 2014) have been stressing out *in defense* of the scan-and-copy hypothesis. It has been advocated that this kind of destructive MU scenario would generate new identities, when actually there is evidence that it would produce the exact same kind of result as the gradual in-placement process, that is, to obtain psychological continuation hence identity preservation. (Wiley, 2018).

Furthermore, it has even been argued that the gradual scenario is technically farther than the scan-and-copy procedure since it would require a kind of molecular nanotechnology yet not available, and eventually not anytime soon. It would probably require a nano-probe or some sort of microscopic neural prosthetic device capable of emulating not only artificial neurons but also their functional connections with each other, which one can envision with the technological pace, but not in the coming years. Yet, conversely, scan-and-copy is in technical reach even almost with the current technology, and as is argued, it would be able to produce the exact same kind of result. (Wiley, 2014; Koene, 2014).

In an eventual nondestructive MU, a duplicate of one's mind could be instantiated, so biological and non-biological minds would coexist, which is quite unintuitive for us, as individuals believing to be one-of-a-kind have a hard time accepting there could be two of us at the same time. This coexistence scenario is more evident in the scan-and-copy hypothesis than in the gradual in-placement procedure, which therefore immediately hinders its intuitive acceptance.

Parfit discusses two (even three) spectrums of identity in his famous *Reasons and Persons* from 1984. *The psychological spectrum* (which follows of a discussion over *the young officer's thought experiment* by Williams), imagines the mixture of two different person's identities but without any biological kind of substitution, resulting in two opposite sets of memory with an intermediate kind that would represent loss and inclusion of memory from both of them. Conversely, Parfit then also imagines the other way around. *The physical spectrum* that would in fact resemble what we today call MU in the gradual in-place replacement kind of version, where the two opposite resulting persons would correspond to before and after material substitution states, with 'the middle intermediate stage', again a kind of *chimera* of some sort. Parfit uses the psychological and physical spectrums to investigate whether identity is closer to a psychological or physical concept [Parfit, 1984].

The question arises of which scenario (if any) would preserve identity, whether the psychological or the physical one. Parfit does investigate how these two transitions of the spectrums should be quantified, although he doesn't find them to be quantifiable, instead he concludes it to be unreasonable to label intermediate points and that there has not actually to be any answer at all! He presents his *Relation R* concept as an alternative proposal to psychological and physical identity. Wiley and Koene argue there is a resemblance in this approach of Parfit's with the kind of reasoning they find to first claim equivalence of scan-and-copy with instantaneous replacement (which corresponds to the intermediate points for Parfit) to then furthermore state the equivalence between instantaneous replacement with slow replacement. This would also dismiss the *all-or-nothing* kind of question of the critique insofar as either one would preserve identity and survive, whereas on the other hand one would either die or have a totally new identity. (Wiley, 2014; Koene, 2014).

Now, just as a reminder before going deeper in this metaphysical equivalence, scan-and-copy pretty much rests on brain preservation techniques and technologies to freeze, section and then scan slices comprising the functional connectome of the original subject in order to build the artificial counterpart a posteriori. In gradual in-placement one would have to develop the nano-probe of some sort, a kind of prosthetic neuron(s) that target specific biological neurons, observe and measure them one-to-one, to later enable their functionality and replace the originals. Wiley's argument states that assuming both procedures would be feasible, thereby producing artificial brains as a continuation of their physical original ones. Hence, exhibiting continuation of their *minds*, there is metaphysical equivalence in both procedures for whatever result there is, it should be identically attributed in both scenarios. (Wiley, 2015).

4.8 _ The Argument of Space (the distance and spatial displacement)

As will be mentioned in the *Critics and Criticism* section, some authors ground their scepticism of a scan-and-copy procedure on spatial and temporal arguments. If minds and their identities would have spatial locations it would then give the impression, in order for a transfer of such nature to occur, that this dislocation of functionality would stand as a barrier. Distance would then by logic be somewhat proportionally related inasmuch as if there were no distance at all, the transferring process would somehow

entail better chances to preserve identity, thus ruling scan-and-copy out and favouring the gradual in-placement scenario. Wiley argues that although originated by the brain, minds are nonphysical therefore also nonspatial and even if this were to seem too dualistic one should keep in mind that at the most it would imply property dualism, certainly not substance dualism. (Wiley, 2015).

But gradual in-place replacement also involves spatial translocation even if the prosthetic neuron to copy the original's *functional connectome* one would not be that far away. Albeit considering it to be an unjustified case because physical distance just does not make sense when approaching a nonphysical function, Wiley troubles himself to argue that ironically the distance involved in MU via gradual in-place replacement is even bigger than in the scan-and-copy scenario, thereby *refuting* that criticism:

“...the micron-scale per-neuron translations that occur during in-place replacement accumulate hundreds or even thousands of kilometers of total discontinuous spatial translation of function, despite the possible presumption that no spatial translation is involved. Once we have already accepted spatial transfer of function on the order of hundreds of kilometers—albeit distributed over billions of neurons—what difference does a few meters across a room really make?” (Wiley, 2015, p7). In summary, and knowing that both procedures would be affected by spatial translation one could determine a kind of threshold in the distance deeming preservation of identity to fail at some point or, as absurd as it sounds, to simply result in another kind of identity, a new one. Alternatively one could accept translations to be independent of physical distance between the biological brain and the prosthetic replacement, hence survival and identity would have to be affected in the same way and the two considered MU scenarios would therefore be functionally and metaphysically equivalent. (Wiley, 2014; Koene, 2014)

4.9 _ The Argument of Time (the rate and temporal displacement)

Is there any difference (in terms of identity preservation) between an instantaneous replacement and a slow one? Also here and again, maybe due to intuitive reasons, it seems that a gradual replacement has gained some ungrounded favour over an instantaneous one which will be argued hereafter. First we have to acknowledge that a gradual replacement, no matter how fast it would be (millions of neurons per second for

the sake of the argument), would always take a tremendous amount of time to complete the procedure, whereas scan-and-copy refers to an instantaneous (and discontinuous) act.

One reason why slow in-place replacement is often favored is that it might maintain some notion of a *stream of conscious continuity* [Chalmers 2010, Van Gulick 2014]). But this is trying to refute an argument with another equally problematic argument. Holding that an eventual stream of consciousness (unproven, strongly opinionated), might be more continuous in a gradual than in an instantaneous scenario, lacks substantiation and is based on mere conjectures. Furthermore, this would then by pure logic mean that sequentially, or consequentially, something should happen along the replacement rate. Because if we were theoretically to speed up the gradual replacement rate, we could reach an asymptote with the instantaneous kind of the scan-and-copy procedure producing the same result. Yet the identity would only be preserved in the gradual kind of uploading and not in the instantaneous one. Does this make sense? Could we then here (*as was the case of the distance in the spatial argument*) also be talking of a threshold of some kind somewhere along that replacement rate?

See, for instance, Wiley: “One might claim that there exists a cutoff along the replacement rate spectrum where personal identity preservation suddenly flips to producing some other identity (despite no change in the physical product).” (Wiley and Koene, 2015, p10). Of course there is a great deal of irony in this statement, but it is not ungrounded and I underline its logical deduction. And so again and in author’s own words: “Immediately below the cutoff, the rate would be considered sufficiently slow to allow total preservation of personal identity, yet at an infinitesimally faster rate, the interpretation would be to utterly reject preservation, resulting in some other identity instead.” (Wiley and Koene, 2015, p10).

One might even pose an *odd question* such as: why would a procedure that in theory produces the same result as the other one (an assumption made at the beginning of this section only to dismiss some of its detractors), end up with a greater proportion of preserved identity just because it took longer? Spectrums of temporal replacement rates, just as of spatial displacements, do not seem to make a difference when comparing the two MU scenarios so until proven otherwise, i.e. in the event of a state-change, metaphysical equivalence in identity preservation should be attributed to both procedures. This rationale aims to either concede that both scenarios are eligible to a

successful MU or, conversely, that neither of them is. So it just doesn't make sense to favour one over the other, at least not metaphysically. (Cerullo, 2013; Koene, 2014; Wiley, 2015).

5 _ Metaphysical Purposes of MU

“For everything begins with consciousness and nothing is worth anything except through it.” (Camus 1961)

5.1 _ Intro

MU has a lot of implications, be it for the definition of who we are as a species and for our future qua species (the discussion on the posthuman), for our own individual self-understanding (the question of personal identity as discussed above), for the understanding of the role technology plays in constantly pushing the boundaries of these definitions and understandings (both individually and for humankind), as well as for the ethical and political qualification of our actions. However, this ultimately boils down to a *metaphysical* question, however elusive and hard to grasp this questioning might be.

What “metaphysics” really is has been a matter of contentious debate ever since Aristotle’s work by the same name came to light with its intriguing title *Ta meta ta phusika*, but it has generally been understood as a synonym of ontology, i.e., the inquiry of “being as such”. Whatever else this might mean, as applied to human reality, the truth is that it involves questions that, as Kant would put it, our reason inevitably raises, even if it won’t be able to solve them – questions concerning, among others, our own future and the fact that we are mortal – or perhaps one day immortal or amortal – beings.

These are of course questionings in which the whole discussion about MU, and the possibilities that it raises, impacts. Namely, and given the – mostly misleading – discussion of MU as a technique devised to achieve immortality, it draws on many “existential” questions. To name just one of the most burning questions on this topic, and one of those that has drawn more objections to MU, we shall have to consider the claim according to which we are inherently finite beings, and that life itself only has meaning if we are finite – something which a MU would allegedly fundamentally alter, insofar as it

could eventually transform us into immortal or amortal beings. Indeed, there is a long tradition of this meditation on the finitude of life and time, from Seneca to Heidegger, and some would even claim that it is exactly finitude that makes life worth living because every moment must be cherished.

Against that backdrop, the aim of this brief section is to spell out the metaphysical implications of MU, by envisioning, also with the help of the thought experiments I put forward, how it should be considered in the light of some existential considerations. Considering how Heidegger defined human beings as *being-towards-death*, and the way in which his reflections on ontology and technology are a source of criticism to the possibility of a MU, it makes sense to start the discussion with Heidegger.

5.2 _ To Perpetuate Consciousness

In many discussions, MU is wrongly linked to extension of life, or better yet - *to the extension of an individual's life*. Many regard this as being the purpose of MU, which as will be herein argued is very much a prejudice caused not only by common popular sci-fi but also because of its endorsement from controversial movements such as transhumanism (and here I am not taking a stand against transhumanism, merely stating a fact). Truth is, MU can be viewed as an eventual mean to overcome natural biological death, but taking this to be its *main purpose* is reductive and even narrow-sighted. Because this selfish kind of purpose is somewhat diminishing of others such as: enhancing intelligence; expanding neurobiological knowledge; triggering philosophical issues to resolve and enabling new ones to burst; curing neurological and psychological diseases; enabling virtual reality with an otherwise biologically unachievable verisimilitude; and of course the main purpose, soon to be argued here, and also because of other reasons which we cannot even begin to think today simply because we still do not have the technology.

That main purpose, and that can be underlined by some of the main actors involved in the actual ongoing developments of MU (as can be seen from/in organizations such as *the carboncopies*), is actually *to perpetuate consciousness*. To enable the possibility of conscious minds to spread through the universe serving an existential concern of humankind is definitely a grander purpose than *an individual's life extension*. “This

approach lends the field of mind uploading — both scientific and philosophical — greater credibility by substituting grand civilizational goals for mere personal goals.” (Wiley, 2018, p1).

One could even argue that MU is *existentially necessary*, in order to provide a fundamental metaphysical purpose not only to human life, giving it meaning, but even to existence, to *Universal Existence*. Existence of conscious beings in the Universe should be perpetuated if only to unravel its true meaning, therefore allowing us to pursue and maybe even answer some of the grand questions that have made philosophers so passionate in their quest.

5.3 _ The Delusion of Reality (in the light of existentialism) - I

The experience of qualia surfing would cause us to pause and reflect on the purpose and meaning of existence. (Loosemore, 2014, p315).

The Delusion of Reality (MU+SI Thought Experiments, Part I, 3c), applies precisely to this purpose. It allows you to reflect on what our consciousness’ purpose is. Maybe it is to develop inasmuch as to seek out for our origins and conversely our fate, giving meaning to an existence that could be questioned in a Universe wide open but full of nothing. “Purpose might even apply to the abstract notion of reality (the fact that anything exists in the first place).” (Wiley, 2018, p2). Consciousness can grant purpose to existence in this sense or even, as explored in the thought experiment, maybe prove reality’s existence, or to conversely acknowledge its virtuality and deem it to be a simulation (consider that there is a lively ongoing debate as to whether our own reality is some form of simulation within an outer reality (Moskowitz 2016). Either way, without consciousness, there neither is reality nor simulation but still existence. And different levels of consciousness could lead us into preserving that existence, again proving its purpose. Furthermore, the thought experiment’s premises forces one to make a choice between being human and just being, without the human. Facing the prospect of giving existence a meaning but at the expense of humankind’s extinction would be a rather unselfish act as one that shows a higher conscious level. The thought experiment can thus be viewed as a clash between an existential affirmation of life and existential nihilism, and as Wiley claims: “When this final conscious event occurs, will the universe,

reality, and existence then cease to have purpose since it will mark the end of the era of consciousness from which purpose could derive?” (Wiley, 2018, p15).

What is life? What is the meaning of life? And what is existence and the meaning of existence in this Universe? Now, one could argue that in order to pursue this metaphysical purpose of seeking answers to these questions we have to increase our knowledge altogether, but more importantly, we need to augment our intelligence and probably its underlying conscious state. Without these tools (intelligence and consciousness) we most certainly will never answer these questions, and perhaps never even be able to pose new questions at a higher level of complexity, which we can only formulate if our intelligence and consciousness evolve to a higher level. We could wait another couple of million years to maybe develop biologically a little more, and eventually *become smarter and more aware* in another few million years – but this is probably more like in another lots of millions of years because biological evolution is very slow. Or, on the other hand, we could *take this matter in our hands and accelerate the process*. We may be able to augment intelligence via Artificial Intelligence (AI) or via Mind Uploading (MU). Either way can lead to the so-called Super-Intelligence (SI) which will be addressed in the next chapter (SuperIntelligence - Part IV). So if not through MU, then with AI. But AI does not entail the possibility of preserving our identities. Nor does MU for that matter, at least not for the time being, but until proven contrary, and this needs to be emphasized, it is *the only technology envisioned to possibly preserve consciousness and personal identity*. AI does not seem to like and eligible candidate for perpetuating humankind’s consciousness and surely not what it is and means to be a *sentient human being*, whereas MU is the best candidate for the job.

Consciousness itself may not necessarily be that fundamental metaphysical purpose, but without it - there is no fundamental purpose to pursue in the first place, hence the importance in perpetuating it. Along that way and just like envisioned in the thought experiment, not only intelligence would be augmented, but also triggered by it - higher conscious levels could develop.

Most (if not all) of the experts interviewed for this thesis showed concern in an existential risk that MU poses to humankind, which is quite easy to deduce from the overwhelming power such a technology would represent, therefore of course also as a potential weapon. The most dangerous extinction risks may be anthropogenic but even then it is arguable that MU as a technology would have been rightfully held responsible since *it is*

not the gun that kills the man but the man who triggers it. Nevertheless, as also some of the experts stressed, despite this particular potential existential risk, the advantages of embarking in such a project as MU outweigh by far these risks and so it would not make sense to simply avoid MU on the grounds of its eventual misuse. It should be noted that *this argument is not against any reasonable caution, but rather against falling into the pitfalls of petrifying fears.* "The best way to thoroughly mitigate existential risks is to spatially spread apart as far as possible. Remaining in a single solar system, much less confined to a single planet, is a recipe for extinction on sufficiently long timescales." (Wiley, 2018, p13).

MU could very well turn out to be the escape route out of a future possible extinction (*here again the thought experiment The Delusion of Reality is applicable*). Since we could be facing extinction scenarios sooner or later on Earth (or even in the solar system for that matter), failing to preserve consciousness, thereby failing to preserve our civilization in a cosmic context, and this makes MU an even stronger candidate for survival rather than of existential risk. (Wiley, 2018). Furthermore, biological interstellar travel may prove less likely when compared to a digital kind of travelling. The biological substance we are made of is not suited for space travel, let alone astronomical distances, which could doom ourselves, letting us trapped inside the solar system in the future, whereas a digital or computerized form would not face these problems. Again, only MU would serve this purpose since it holds open the possibility of preserving our personal identity as opposed to AI (or even AGI). (Wiley, 2014). From the above we could then defend that maintaining consciousness via MU would actually be a way to avoid extinction, which one could now argue to be a very vague proposition since extinction could either never happen or only in a distant future. If the probability of a non-anthropogenic nature (e.g. supernovas, black holes, solar explosions, comets etc...) kind of disaster happening is so low - then it would not make much sense to bother with it or at least not for now. But is that probability that low? If we consider recent viral pandemic outbursts to have proven our vulnerability and shaken all humankind in every possible structure quite a lot, then humankind's glorious future suddenly does not seem such a safe bet afterall. What if the next virus would be just a bit more virulent or contagious?

Anthropogenic risks (natural disasters) are not so improbable but this alone does not constitute an argument to delve into what could turn out to be the most expensive project in humankind's history. Yet, there are those that evoke existentialism in order to sustain

the attempt to preserve and perpetuate consciousness in space and time, and I will engage in this discussion, following the thought experiment's (*the delusion of reality*) premises, as this may constitute one of MU's major objectives.

5.4 _ The Delusion of Reality (in the light of existentialism) - II

I am condemned to exist forever beyond my essence, beyond the causes and motives of my act. I am condemned to be free. Jean-Paul Sartre (*Being and Nothingness*, (Sartre, 1956,p.438-9)

Keeping *the delusion of reality* in mind, and drawing some insights from philosophers such as Husserl, Heidegger, Sartre and Camus, this small section is an incursion into phenomenology and existentialism with the help of some of these great thinkers. It is not meant to be a pretentious pseudo-analysis of their works, the goal rather being to highlight the thought experiment's premises and its provocative outcomes given this background. Thus, in the light of existentialism and using the paper "*Integrative Perspective Martin Heidegger and the question of being*" by Emmanuel Iwuagwu, as a backbone to analyse *the delusion of reality* as a thought experiment, and exploring the purpose of MU as an envisioned technology to perpetuate consciousness and to seek *meaning in existence*. We will argue that reality is what we experience it to be.

Heidegger's primary concern was ontology and although he did not see himself as an existentialist, his *Being and Time* is seen as a reference in existentialism. *Being and Time* starts with an ontological question: "what is being?" What does the verb to be actually mean? Heidegger formulated this as *Seinsfrage* (question of being). To answer this question, Heidegger differentiates between what it is for beings to be (*Sein*) from what it is to be for Entities in general (*Seindes*) to then focus on the being that is engaged in the world (*Dasein*), the being that understands what it means to be and that questions it. It is useful to recall that for Husserl, the primary goal of phenomenology consists in its descriptive, detached analysis of consciousness in which objects, as its correlates are constituted. For Heidegger, in contrast, the starting point of philosophy is *Dasein* in its being and not consciousness. On the other hand, for Husserl, the problem of constitution is the core problem of his philosophy. For Husserl, the main question is: how is the world as a phenomenon constituted in our consciousness? According to Heidegger, Man

differs ontologically from all other existents. Only Man can pose the question “what is the being of beings?” and therefore only him can make an ontology possible. It is this ability to pose this question that leads man to transcend his own being.

As Iwuagwu makes clear, “according to Heidegger, *Dasein* owns his own being. He has responsibility for his own being, without being responsible for being there. Hence whereas other entities merely are, *Dasein* exists. *Dasein* is synonymous with *existenz* but *existenz* is not some fixed quality or essence. On the contrary *existenz* is a constant possibility (1956, 124).” (Iwuagwu, 2017, p25-48).

One of the consequences of Heidegger’s “existential-temporal analytic”, and of his critique of onto-theology (what he sometimes calls the “metaphysics of presence”) is the radical diagnosis of the absence of all pre-given foundation to existence and hence, at least as it was interpreted by later existentialists, the need for human beings to provide themselves meaning for their own existence – a meaning that does not exist by itself, outside of our capacity to provide meaning. If *existenz* is a constant possibility, and one that depends upon *Dasein*’s responsibility in giving meaning to it, in upgrading from merely being to actually being there, existing - then in the TE we can relate this to humankind’s responsibility in using *the glitch*, using the possibility in eventually giving meaning to *being* by undergoing the restart.

According to Heidegger, “Existence is the specifically human mode of being, the distinguishing characteristics of *Dasein*. All other beings merely are; *Dasein* alone exists.... *Dasein* is not merely something there, not a pure object like a rock on the beach. Rather it has responsibility for its own being and an obligation to be significant. *Dasein* recognizes its own being as a task to be fulfilled; it must do something about its being” (1956, 124; Demske 1970, 16). In *Being and Time*, Heidegger insists the essence of *Dasein* lies in its existence”. (Iwuagwu, 2017, p25-48).

In *the delusion of reality* we can translate this reasoning by highlighting the obligation to be significant and not just to merely be. It would make no difference whether we would be in reality or in a simulation since it is the *Dasein* that actually defines the reality of that world.

Also, at the end of the TE the options are laid down and humankind is forced to make choices but it has in its power the means to do something about *being*.

This is what Heidegger defines as resoluteness and authenticity, in that, in choosing, *Dasein* also decides its way of life, i.e., falling prey to an inauthentic going about and doing “like everyone else” (*Das Man*) or disclosing the truth about Being and about oneself by choosing to live authentically. As Iwuagwu makes clear, “And because *Dasein* is in each case essentially its own possibility, it can in its very Being, “Choose” itself and win itself.... In determining itself as an entity, *Dasein* always does so in the light of a possibility which it is itself and which, in its very Being, it somehow understands. This is the formal meaning of *Dasein*’s existential constitution” (1962, 67 – 69).” (Iwuagwu, 2017, p25-48).

Heidegger’s rationale tempts me to make a reasonable inference: According to *the delusion of reality* thought experiment, and in light of the *Dasein*’s existential constitution, could humankind face the options evoked in that TE and not choose *to be*? Even if in that particular case this would have meant cease to be as human but allow it to *really be*? Because in that case (*superconsciousness level hypothesized*) they would have understood what it means to be as *Existenz*, and to make authentic decisions according to a higher understanding of reality.

Since *Dasein* is his own possibility he is never something completed, stable or fixed. *Dasein* never quite is but is constantly realizing his possibilities through his existential choices. The most basic state of *Dasein*, being-in-the-world, means human existence is immersed in the world of actually existing concrete individuals. Man is inconceivable in isolation from the world and neither is the world meaningful without *Dasein*. (Iwuagwu, 2017).

This, in the delusion of reality scenario, reinforces the simulation as *a possible reality* as long as it gives meaning or significance in the relationship of the *Dasein* with that *new reality*. Both make up one reality, and in that TE, humankind (*Dasein in the TE*) realizes precisely that, labeling the endless simulations state as unreal, because there can be an upgraded *Dasein* in another environment that does not force Man to tie up with it in order to reel it.

Heidegger then devoted some 130 pages of *Being and Time* to analyze separately the notions of ‘Being-in’, ‘the world’ or the ‘self’ who is in the world. Being-in, far from designating a spatial relationship, indicates an existential relationship between *Dasein*

who is the Being-in-the-world. Thus, without the world *Dasein* can neither be nor be understood. (Iwuagwu, 2017).

Finally, translating into *the delusion of reality* context, we might assume that a new level of consciousness (achieved via MU) could empower humankind in its search for authenticity in reality (whether a fabricated simulation kind of reality or not), therefore also get closer to *existenz – whatever it should mean*. In the TE, providing that reality, which I compare to Heidegger's *Dasein*, could enable humankind to give meaning to existence and life itself. Heidegger's rationale in search for the being and in search for the being giving meaning to its existence, finds some parallel in the TE's rationale of the same sort of metaphysical quest, hence justifies one of MU's principal objectives.

5.5 _ Essence and Existence

As opposed to Husserl's phenomenology, consciousness is not the core, rather a possible mode of the *Dasein's* being in Heidegger's philosophy. Contrary to a consciousness' descriptive analysis, to Heidegger, it is the *Dasein's* being (to be there) that can unravel the transcendent constitution of the world. As Sartre and Kierkegaard, also Heidegger was convinced that a physical body preceded the essence of self. To Heidegger, a being becomes aware that it exists at some point of its developmental process. This is the moment where essence begins to shape or the moment when a subject acknowledges and embraces essence.

"The 'essence' of Dasein lies in its existence." (Heidegger, *Being and Time*, p. 67, 1964). The "essence" of this entity lies in its to-be [*Zu-sein*]. Its being-what-it-is (*essentia*) must, so far as we can speak of it at all, be conceived in terms of its being (*existentia*). (Heidegger, *Being and Time*, p. 67; H, p. 42). This was an inspiration to Jean-Paul Sartre. Sartre's *L-Être et le Néant (Being and Nothingness)* follows in the line of Heidegger's *Being and Time* as a main influence as can be seen in many aspects of his thinking like his position regarding the essence of man, which he also believes to precede that historic existence which we confront in experience. (Blattner, 2017). But there are of course also noticeable differences between Sartre and Heidegger, especially if we take into account the influence of Nietzsche's claim on the death of God, Hegel's *Phänomenologie des*

Geistes, as well as Marx's dialectical materialism (among others, of course) in Sartre's existentialism:

"If man as the existentialist sees him is not definable, it is because to begin with he is nothing. He will not be anything until later, and then he will be what he makes of himself. . . . Man simply is. Not that he is simply what he conceives himself to be, but he is what he wills Man is nothing else but that which he makes of himself. That is the first principle of existentialism. . . . Before that projection of the self nothing exists Man is responsible for what he is. Thus, the first effect of existentialism is that it puts every man in possession of himself as he is, and places the entire responsibility for his existence squarely upon his own shoulders" ("Existentialism Is a Humanism," included in *Existentialism from Dostoevsky to Sartre*, ed. Walter Kaufmann, pp. 290).

"Atheistic existentialism, of which I am a representative, declares with greater consistency that if God does not exist there is at least one being whose existence comes before its essence, a being which exists before it can be defined by any conception of it. That being is man or, as Heidegger has it, the human reality. What do we mean by saying that existence precedes essence? We mean that man first of all exists, encounters himself, surges up in the world — and defines himself afterwards." ("Existentialism is a Humanism," in *Existentialism from Dostoevsky to Sartre*, ed. Walter Kaufmann, pp.289-292).

Now, Keith Wiley extends Sartre's claim beyond the individual and extrapolates it to a fundamental purpose at the level of the Universe and so states that in this line of reasoning, Sartre could say that the Universe's existence precedes its essence. "If consciousness underlies purpose, the universe and reality are then necessarily born without purpose since at the outset they lack conscious members. With the evolutionary arrival of consciousness, the universe then moves to the next stage, in which its purpose is discovered or created from within." (Wiley, 2018, p15).

Following this rationale, as *conscious beings thrown into this Universe*, we are then granted to provide it with a purpose, the purpose of reality and existence, almost as if we were responsible for it. But we could also approach this holding the premise that there is no fundamental purpose, as would have been Albert Camus' standing. In Camus' reasoning there is no fundamental meaning of life and his solution to that lies in our acceptance of it, which he called the absurd realization (*philosophy of the absurd*, Camus 1961). Son the one hand, to embrace it as absurd truth would be part of his solution.

Accepting an undesirable truth over the preferable alternative of irrational denial and self-delusion, which in the case of the thought experiment corresponds to the acceptance in the cycle of simulations that reality is made of. On the other hand, also part of Camus' solution would be to revolt against it. To accept cosmic futility while refusing to accept the expected despair, which corresponds to humankind's choice in starting-over evolution in order to give existence an opportunity.

6 – Critics and Criticism

Given that MU represents a controversial technology in so many ways, it is therefore under continuous scrutiny whether from experts, stakeholders, people from science, people from technologies, philosophers, people from religion, people from politics and so forth. Since it has been explored in sci-fi for many years and discussed in a philosophical environment for many others, its criticism generally took a philosophical language and was also generally debated in that field. But in these last couple of decades (namely since the beginning of the 2000's), because of a great deal of progress in the related fields of NBIC, *nano-bio-info-cognitive*, (neurosciences, nanobiotechnologies, artificial intelligence and even whole brain emulation as an emergent technological field), its scrutiny has become very much technological. Nevertheless, the main field deriving criticism generally speaking upon MU remains grounded on ethical and other related issues.

Some of the main usual kinds of criticisms exercised on MU are herein described, along with *my* correspondent counter-arguments. As an analogy with *frequently asked questions* (FAQ), I will designate these disclaimers as *frequently made objections* (FMO). Some of them were specifically made and used in recently published papers and their author's claims are herein presented as being examples reflecting the position of many other authors.

6.1 _ Frequently Made Objections

6.1.1 _ *Only quantum computing would be able to emulate neural processes (hereby dismissing "regular computing power")*.

Roger Penrose has been one of the leading scientists in this quest of attributing quantum mechanics to consciousness processes, mainly in the nineties (Penrose, 1994). But until now there is no evident data supporting this idea, rather the contrary and contra Penrose and Hameroff no crucial role in how the brain processes sensory information, stores and retrieves memory or generates behavior has been linked to quantum mechanics. (Wiley, 2018). Nevertheless, one can always argue that quantum mechanics has since the nineties come a long way, namely with quantum computing, which ironically could turn out to be not an obstacle in MU project's path, rather a catalyst. Only very recently some great breakthroughs have been reached in this field showing that at least its progress rate holds big hopes in contributing to build prosthetic brains.

6.1.2 _ *MU is such a large undertaking it deems itself unachievable.*

It is true that MU is a large undertaking (Bonilla, 2017) but to deem it impossible is like all historical claims deeming impossible all those scientific and technological milestones humankind then later proved to be surpassable. Yet there are sceptics who judge computing's challenge a kind of near-impossibility task to match the brain's scale, although there is no such evidence when regarding the pace of technological progress. "The mind is an informational process precisely because its transfer functions can be achieved computationally." (Wiley, 2019, p1). Also, if this kind of argument would always apply then man would have not walked the moon yet.

6.1.3 _ *We lack a kind of "neural Rosetta code" which blocks us from ever understanding how the brain processes information. It is like trying to decode a CD without having the CD player's software (Bonilla, 2017).*

This line of reasoning underlines the idea of the brain being a kind of black box, but in fact, even the most opaque information processing system can be deciphered. This is precisely the reason many neuroscientists (Koene, Deca, etc...) approach the issue in terms of system identification in which all inner details are abstracted away, leaving only their input and output data and the inferred transfer functions, as has been already sufficiently described in the previous WBE chapter. There actually is no missing information, the data is there to be uncovered and the approach we use in trying to decipher it, is what drives the ongoing projects towards its ultimate decoding.

Furthermore, as soon as barriers are overcome, new ones arise leading sceptics and general critics to claim it impossible yet again. “Detractors of the you-know-nothing-if-don’t-know-everything persuasion can always move this goal post down one level... At whatever level we draw that line, we are black-boxing the underlying mechanisms from that point on. Thankfully, the determination of the correct level of abstraction needn’t be ad hoc; it can be achieved experimentally.” (Wiley, 2019, p1).

6.1.4 _ *Minds only operate properly in the context of an associated body and its sensory-motor interactions with the physical world.* (Lakoff, 1980).

The so-called embodied cognition argument fails at a first stance by the simple fact that MU could be carried out also in a scenario involving an embodiment of the prosthetic brain. However, this is actually a metaphysical question only answerable after MU would have been accomplished, and in both kinds of scenarios, with and without embodiment. Only after this, one can (or could) judge on its absolute importance as a requisite of a successful MU. We won’t be able to do this before a MU is achieved, unless some kind of simulations could be carried out, which ultimately would not have the same meaning as an actual emulation. Nevertheless, there are a variety of scenarios envisaged concerning embodiment, whether virtual or not, only finding an analogy in the actual medical condition of *locked-in syndrome patients*. (Wiley, 2018). In these cases, it is not impossible to carry out a life disconnected from sensory and motor access as long as the mental health allows it. But to try to establish some real comparison in these particular cases is not a safe analogy for there are also psychological factors that make these two scenarios differ quite a lot. Anyway, and as stated, embodiment could be an option along with MU, but it would not affect its technical feasibility, meaning MU could be achieved even without it.

Sandberg has also claimed that it would make sense for a project of WBE and mind uploading to be realistic that it should include a body model (and this even though WBE and mind uploading refer to the brain, to the mind), to allow the emulated nervous system to best interact with a simulated world (or even with the real world, for that matter). He also makes reference to the physiological feedback loops that might influence neural activity that is part of a body-brain dynamics. (Sandberg, 2014).

Mind uploading would change the nature of embodiment and according to Linssen et al it would not abolish it. It is the body that mediates interactions with the environment, and the brain is in fact linked to this mediating interacting body. In fact, our brain is the way it is, and has its characteristics strongly influenced because of the relation it has to the body and its surrounding environment. MU would ultimately mean changing the host (body) to another substrate but it would not eliminate that relation. We might research and investigate how really to cover this relation, but surely it would have to be compensated – as millions of years cannot simply be wiped out without leaving an impact. There is an *embeddedness* that refers to the environment in which our bodies (containing our brains) fit. After uploading, the transition does not necessarily alter our perceptions of that environment or of our body. This *embeddedness* (see Figure 22), allows us to interact with the phenomenal world *out there*. The body is subjected to the same physical laws and constraints of the world it relates to and constitutes the bridge to it, as Linssen et al claim: “Thus, in an uploaded existence, the perceived embodiment can be vastly different from the substrate. Embodiment in WBE should thus refer to the body and environment that we perceive to be physically real, even though it may not actually exist in the real world”. (Linssen et al, 2016, p4).

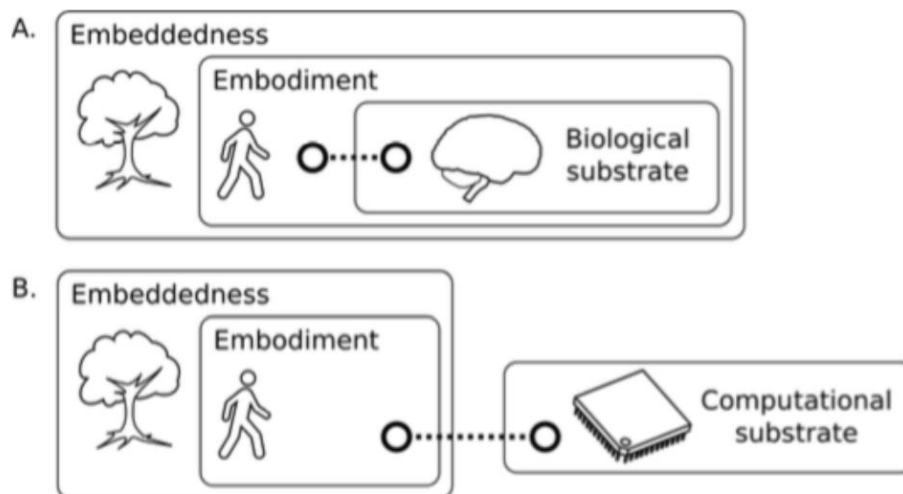


Figure 22 (adapted from Linssen et al): Diagram relating embeddedness, embodiment, and substrate. (A) Present situation: the brain is contained in the body which is contained in the environment. Action potentials are routed between brain and body via nerve fibers (dashed line). (B) Situation for WBE: the substrate is no longer fixed to the body. Action potentials are electronically routed between the emulation and body (dashed line). (Source: Linssen et al, 2016).

Furthermore, Linssen et al argue that one cannot really distinguish “pure” sensory processing from “pure” higher cognition. In their rationale, they find support in the fact that development of symbolic thought such as semantics in language, rationality and imagination, is fundamentally dependent upon the body’s interaction with its surrounding environment. They evoke several examples showing that higher-level concepts are metaphors that ultimately rely on our bodily experience of interacting with the world including our relation with time. They claim that our interaction with the world through seeing and moving ultimately shapes our understanding of the past and the future and that these abstract concepts we form in this relation would mean nothing to us without a basic sensorimotor experience. (Linssen et al, 2016). To summarize their claim: “it is misleading to refer to areas of the brain as if they were modular. Instead, it is impossible to draw strict boundaries separating, for example, the more primitive sensorimotor functions from higher cognition such as planning, judgment, decision-making, and direction of attention. Furthermore, these areas are highly integrated, functionally and anatomically, in a complex and dense network. Thus, the “quintessentially human” functions cannot exist independently of brain structures that are devoted to world interaction and body control. Even for an uploaded individual, a body (human-like or otherwise) would remain a necessity.” (Linssen et al, 2016, p6).

An instantiation in a virtual environment of emulation of the brain but also of the body and the environment itself might work as an alternative to a real-world avatar. This is actually the most common scenario envisaged that could be a feasible strategy to compensate the lack of a biological body. The human brain is quite flexible and able to quickly and effortlessly adapt to new types of embodiment and to learn new capabilities and to learn to control them. (Pino et al. 2014, Linssen et al, 2016). Thus, it seems reasonable to infer that the embodiment issue has viable solutions already thought of.

6.1.5 _ *There is no reason to suppose an exact functional copy of the brain will produce similar phenomenological effects and also that the uploaded mind cannot entail the same self as the original brain.*
 (Hauskeller, 2012).

My first intuitive answer to this claim would be *there is also no reason to suppose an exact functional copy of the brain will not produce similar phenomenological effects* and so on.

But, objectively, Hauskeller (among others) finds it difficult if not impossible for an emulation to elicit the same functional brain as the original one, let alone its underlying personal traits (self). But he nevertheless concedes, and this might be possible soon, to replace parts of the brain (not as enhancement but for therapeutic reasons): “Even the brain itself might be replaceable. Alzheimer patients and others suffering from the effects of brain damage located in the hippocampus may soon be helped by having the damaged parts replaced by an artificial brain prosthesis, a microchip hippocampus [Berger, 2005; Berger & Glanzman, 2005].” (Hauskeller, 2012, p2). However he then directs his critique to supposedly dubious motifs behind these enhancements or therapeutics, and in this way discloses what strikes me as being his real critique, based on moral grounds: “Soon we could be meshing our brains to computers, living, for all practical purposes, on an ‘immortal’ substrate, perhaps eventually discarding our messy, aging, flesh-and-bones body altogether.” (Klein, 2003, cit. by Hauskeller, p2). The human body is not only regarded as dispensable; it is an obstacle, an enemy to be fought and to get rid of. It ages and makes us age with it, eventually annihilating us.” (Hauskeller, 2012, p2).

As presented earlier in this thread, the main metaphysical purpose of MU is to technologically allow the perpetuation of consciousness, which most certainly has noble motifs (if not only to avoid extinction) thus providing MU with a valid purpose beyond that *selfish and morally arguable individual life extension motive*. Furthermore, and even admitting as viable the argument according to which our biology to some extent is a barrier in the perspective of humankind’s progress and development as a competitive species (even in space), to regard the body and biology as *an enemy that has to be fought and get rid of*, is clearly a dramatic wrong choice of words. It doesn’t even make much sense to feel such an abjection of one’s own embodiment or biology nor does any of the proponents address our biology as such.

“Whether or not ‘the phenomenological effects of a mind’ will indeed appear obviously depends on what the mind is and in what way it is dependent on its existence on the body. We would have to assume that some form of functionalism is true, that, in the words of AI pioneer Marvin Minsky, minds are what brains do [Minsky, 1986: 287]. This generally means that the mind supervenes on the physical and is not dependent on the specific material constitution of what it supervenes on. We are to assume that the mind

is based on the functional relations between physical elements and not on those elements themselves.” (Hauskeller, 2012, p2).

Hauskeller, as others would in this case, does concede that until an actual WBE would have been carried out, we cannot simply speculate on something we still do not even know, which I would consider to be an argument against himself (themselves). If we still cannot define accurately what a mind means, what it entails and to what extent its information is or isn't reproducible, then arguing against its feasibility does seem as walking on thin ice. Moreover, he re-directs his discourse toward the embodiment issue, which as addressed earlier, is not an issue regarding feasibility.

6.1.6 _ But even if the mind resulting from a particular brain emulation will indeed be qualitatively identical to the mind it is supposed to copy, the two minds also need to be numerically identical, which is an additional requirement. (Hauskeller, 2012).

I do not think (as has been clearly and sufficiently meditated by MU proponents), that in order for an emulation to succeed in a functional upload, it would have to result in numerical identity. I do believe this to be a romantic idea (or maybe not romantic at all, rather *quite scary*) many people have regarding MU, that there would have to be two equal beings living at the same time. As was also clearly stressed by the majority (if not all) experts interviewed for this thesis, an upload would *at best* be equal to its original up until that moment of conception (upload) and from there on experience its own *stream of consciousness*, thereby distantiating itself from its original in time, hence *building its own self*.

However, Hauskeller insists in this mathematical issue of two things equal, not being actually equal rather only alike: “This means that for the successfully instantiated mind to be mine, it would not be sufficient if it were in every respect indistinguishable from my mind. Rather, it would have to be literally the same. There seems to be a conceptual difference between a mind that actually is mine, and a mind that merely thinks, feels, and remembers exactly as I do, and hence is like me in all respects, except that it happens not to be me. Think of two copies of the same book that, despite having exactly the same design and content, are still two copies and not identical to each other.” (Hauskeller, 2012, p3).

I will take his analogy with the book and counter-argue that the book would entail the exact same content of its original one, and in its absence would represent it in the best possible way. Now, this would not mean it would be the same book, as also argued earlier in this thesis, numerical identity, is in fact a mathematical impossibility. But this would not diminish it in any aspect whatsoever, and this is the actual twist: It would not make a difference in the end, as long as one's consciousness continuity would have been assured, if it from there on would experience its own path inasmuch as I do not expect to be exactly the same in ten seconds, for I too, am a subject of the passage of time and everything it carries with it.

Moreover, and in respect to the numerical identity required by the sceptic: to obtain a personal identity, it would correspond to a mathematical identity, which essentially instantiates that two identical states between common-pared objects which would then also exclude anybody who has lived more than about twenty years or so, since anyone's childhood *self* has also not the same identity as its middle-aged counterpart, let alone old-aged counterpart.

6.1.7 _ *“The ancient paradox known as the Ship of Theseus, in which we are asked to imagine a ship that is being maintained by having those of its parts replaced that are no longer functional, until eventually all of its original parts are gone and everything is new. The question then arises whether the ship is still the same ship or rather a different one.”* (Hauskeller, 2012, p5).

This is a frequent analogy to the gradual replacement MU procedure in the sense that in that classic ancient paradox a ship's parts are replaced gradually and at the end, the ship although being entirely new, would have always been the same. Or maybe not. That is the paradox, and it is widely used to show that contrary to ships, humans have identities that could not be reproduced simply by changing the brain's components, as in the ship's analogy. It is argued that the graduality process should be considered fallacious because it supposedly denies the reality of change. “The self may survive the gradual cyborgization of the human body through various changes, but only up to a certain critical point. When that point is reached the self might disappear (to be either

replaced by a different self or to vanish entirely without being replaced at all).“ (Hauskeller, 2012, p6).

The counterargument comes from another paradox, known as *the Sorites paradox*. If we start removing grains of sand out of a heap of sand, until we finally are left with only one grain of sand, do we still call it a heap of sand? At what moment, at what particular grain of sand did it cease to be a heap, and similarly at what point in gradual replacement, did the transfer occur or at what point did the person we used to know cease to be as it was? Hauskeller (and others) argue that there is no definite threshold where one becomes another person. Again, using the Sorites paradox as a reference, there is a spectrum between the endpoints starting from the removal of that first grain of sand until the last one. But in this case we clearly consider two different endpoints. One that has all the grains of sand which will unambiguously be called a heap of sand, and one made out of only one grain which we may find a hard time in calling a heap. So the difficulty here lies in determining the exact point (or grain of sand) corresponding to that transition along the spectrum, or in the case of the identity transfer at what point it would actually take place. Here, proponents of MU would argue that the question is not when or where that identity transition is supposed to occur, but whether there is a difference between the endpoints to begin with! If there is no difference, then the paradox cannot even be applied. (Wiley, 2018). As analyzed earlier in reference to the space and time issues in an uploading scenario, we *still cannot* claim identity to be dependent on translation distance and therefore we should by default adopt the simplest theory which is the functionalist view.

6.1.8 _ *Information idealism and the walking algorithms argument.*

Hauskeller argues that some form of *information idealism* reduces us (human beings) to a kind of *walking algorithms* in the sense that, as he claims it, many contemporary biologists as Richard Dawkins label living organisms as vehicles of preservation and transmission of information. Although he even praises Katherine Hayles in stating “Underlying the idea of cyberspace is a fundamental shift in the premises of what constitutes reality. Reality is considered to be formed not primarily from matter or energy but from information. Although information can be carried by matter or energy, it remains distinct from them. Properly speaking, it is a pattern rather than a presence.” [Hayles, 1996: 112]. Despite his admitting of a kind of metaphysical stance in this idealism, he

rejects the idea of human beings being these *walking algorithms made of bits of information* and targets his critique, not to information or transmission of information itself but information processing: “Having a mind generally means being to some extent aware of the world and oneself, and this awareness is not itself information. Rather, it is a particular way in which information is processed.” (Hauskeller, 2012, p7).

From his rationale it seems a reasonable inference to make that the complexity involved in our brains when processing information is what could not simply be transferable, whereas information per se could. But this complexity, this processing of information is what MU proponents call the pattern of information, and by pattern they mean not only the information per se but its complexity too. As addressed earlier, consciousness would for a functionalist be made of these patterns, which, as organizational invariants, are not bound to the physical matter they were to be instantiated.

6.1.9 _ “*Although, let’s face it, we don’t have the slightest clue how conscious experience comes about and how there can be such things as selves in the first place, it is rather unlikely that mind and self are directly produced by the brain, as is commonly assumed.*” (Hauskeller, 2012, p7).

This is surely an interesting claim. Although one admits not knowing the process that gives rise to consciousness, one can nevertheless be quite certain of the reverse stance, that is how it cannot occur. Though it seems quite controversial and even somewhat absurd, it is not that infrequent to simply claim the impossibility of something without even knowing what that something is in the first place (hence its inclusion in this section). It is commonly used along with the also highly opinionated argument that the transfer of such things as consciousness and identity cannot be simply identifiable or even processed because they seem too complex. Usually they tend to end up with the further facts kind of reasoning entailing the soul argument, which as mentioned before, I cannot address for being too subjective and out of the scope of this thesis.

6.1.10 _ *Regarding instantaneous destructive uploading, a recurrent objection has to do with the supposed distance to travel once the*

information would be transmitted from the original brain to a computer. (Corabi & Schneider, 2012; Corabi, 2014).

“Consider first that the mechanism by which the person (i.e., the propertied substratum) would move instantaneously from the brain to the computer is problematic, even on the assumption that only a short distance needs to be traversed. Not only does this involve an unprecedentedly rapid kind of motion for a person to follow, but this sort of motion is oddly discontinuous. For it is not as though the person moves, little by little, to the computer, so that a step-by-step spatial transition from brain to computer can be traced” (Corabi & Schneider 2012, p10).

The counterargument lies in functionalism holding that by definition minds are nonphysical therefore also nonspatial. Functionality does not move through space *from A to B*, and thus it should not matter whether the distance would be small or big. (Hopkins, 2012; Wiley, 2015?). “In fact, we advocate for the stance that function, mind, and identity don’t even have spatial locations to begin with and that any concern or challenge about their alleged motion through space is a misnomer.” (Koene & Wiley, 2014, p9).

6.1.11 _ “*The main reasons for being pessimistic have to do with reservations about positions that attribute strange and discontinuous motions of persons (in contrast to the behavior of other macroscopic substances) in order to claim that uploading is compatible with survival.*” (Corabi & Schneider, 2012, p13)

“The proposal of a cutoff or smooth blend in spatial translation requires that some metaphysical property of personal identity can tolerate spatially discontinuous translation over a distance of tens of microns, but seemingly not over a distance of centimeters, meters or perhaps kilometers—or light-years (including vast accumulations of distance gathered from billions of otherwise micron-scale translations in the case of in-place replacement, as described above). There is no evidence that such a metaphysical property of consciousness and personal identity survival exists that is subject to these seemingly bizarre physical and spatial effects, especially considering that we are

assuming perfect technical efficacy (and, in fact, identical physical results) across any such spectrum of spatial translations.” (Wiley, & Koene, 2015, p9).

“The most rational conclusion to the questions that arise when considering spatial translations of identity is that identity and survival simply have no relationship to the spatial distance between the biological brain and the prosthetic replacement.” (Wiley & Koene, 2015, p9).

6.1.12 _ *Regarding gradual destructive uploading, a recurrent objection has to do with temporal discontinuity.*

“...does the person’s substratum instantly (or almost instantly) travel from brain to computer? Is there a temporal discontinuity between the person’s disappearance and reappearance? (Corabi, 2014, p14).

“...even dualist views that do not require physical building blocks to be preserved would require the physical bundles that are supposed to causally interact with the nonphysical bundle (i.e., the soul or immaterial mind) to endure in spite of spatiotemporal interruption, and to causally interact with completely different physical properties before and after the temporal break.” Corabi, 2014, p14).

Corabi and Schneider claim odd spatiotemporal discontinuities to be their cause for disbelief in both MU procedures, mainly in the instantaneous uploading scenario. These were already addressed earlier in the previous section (metaphysical equivalence in uploading scenarios) of this thread. To sum it up, see Wiley: “As with spatial displacement, any purported distinction between various rates of replacement proves to be highly problematic, and furthermore, to be metaphysically unmotivated in that (exactly as was shown for spatial considerations) there is no particular reason to initially believe identity bears a relationship to replacement rate in the first place. Consequently, until shown otherwise, the more parsimonious and more likely conclusion is that there simply is no such relationship.” (Wiley & Koene, 2015, p10).

6.1.13 _ *“I combine the claim that there is a significant chance that mind-uploading will fail with the claim that comparatively little is gained if the process is successful and much is lost if it fails. Hence,*

my conclusion that mind-uploading is prudentially irrational” (Agar, 2011, p432).

Philosopher Nicholas Agar was criticized by MU proponents for claiming that MU is *prudentially irrational*, which made him write a paper arguing his statement. One of his major premises relies on what he calls to be the “Searle’s Wager” and the possibility of the so-called strong AI failing to deliver, which ultimately could lead MU to be unsuccessful. Searle’s Wager imagines candidates for mind-uploading being asked to place a bet. The success of mind-uploading is contingent on the truth of Strong AI. If Strong AI is a correct view then the procedure may work but if not then MU would fail and, in the eventual destructive scenario envisaged - *kill* the original person in the process. Neil Levy (one of the proponents of MU) made a *reductio ad absurdum* out of Agar’s reasoning: He acknowledges a non-zero probability of *Searle’s Wager* in being right and thereby proving MU to be fatal, but he also acknowledges a non-zero probability in gene manipulation being fatal as there actually are lots of other non-zero probabilities *of things going wrong and being fatal*. Thus, rejecting MU because of a non-zero probability would hardly be reasonable.

Agar defends himself showing probabilistic motifs to be relevant when balancing the pros and the cons of such a technology. Also interviewed for this thesis, Nicholas Agar used in that same interview an analogy to underline his position: Crossing a street has an associated risk of death and the benefit of getting on the other side of the road may prove to be worthwhile taking that risk. Granted, it is a low and balanced kind of risk, and that is exactly what one should do about MU’s risks. To balance them before eventually undergoing such a procedure. We might even find those risks acceptable when counterbalanced by the possible benefits of a successful MU. Moreover, Agar’s principal claim lies on *the principle of epistemic modesty* which he recommends Strong AI proponents should accept. The principle basically states that one should accept the possibility of being wrong and opponents being right. He furthermore argues that many of the mysteries (as he calls them) about thought and consciousness will still be there even with an intelligence augmentation that could come with MU. I would counter argue that precisely many of those mysteries could be unveiled with MU and augmented intelligence. Agar agrees with Levy inasmuch as a failure in MU would not necessarily be fatal and that it might result in some kind of *mindless survival*. However in the end it is Agar’s opinion that such a procedure would present little to gain and more to lose.

To such cases of balanced risks ending up with supposedly more risks than benefits I would simply counter argue that landing on the Moon had a lot of risks. It was worthwhile. It still is. Exploring Mars may have a lot of risks. But it is worthwhile. Quoting Ken Hayworth from the interview he gave to this thesis: “The idea behind being scared of this future and that we’re going to lock ourselves away from a cure to all diseases including mental diseases; From a range of experiences that biological evolution simply hasn’t given us access to; From space travel etc... Those advantages have to be sacked up against the scary aspects.” (Hayworth).

6.1.14 _ *“Consciousness is - to the best of our knowledge - a biological phenomenon that is unlikely to be entirely (or even largely) substrate-independent.”* (Pigliucci, 2014, p119).

Massimo Pigliucci made a relatively short, yet very concise and incisive counter-analysis to David Chalmers’ philosophical analysis on MU first presented in the *Journal of Consciousness* in 2010. Pigliucci himself acknowledges Chalmers’ arguments to be one of the best defenses of MU, which is what led him to try to deconstruct Chalmers’ reasoning. Hence his conviction titling this section although he does concede that consciousness could be generated in substrates other than the current ones. This seems a little ambiguous but actually underlines Pigliucci’s relevance in the substrate issue. To begin with, his criticism focuses on the vulnerability of the CToM (Computational Theory of the Mind), which he argues to be an incomplete theory, but that it is one of the MU proponents’ main arguments for consciousness generation independent of the substrate. One might counter that Einstein's Relativity theory isn't complete or that Beethoven's ninth symphony is unfinished, yet that does not make them less true. As previously stated in this thread, Pigliucci’s view on MU is that it would stand for a mental cloning, a duplication at best, but never for consciousness preservation.

Pigliucci starts his criticism of Chalmers accusing him of acknowledging we still do know little about consciousness and that this alone should stop him from proceeding. I think this could go either way. Because one does not know how the universe starts does not mean one cannot presume it could have started in a particular manner which then took us here; or even precisely because one does not know exactly how consciousness is generated by the brain we could presume it could be generated through other media besides the biological one, which is to claim: it could be generated in other substrates

because we don't have any proof that it couldn't. I guess we can conclude that even though this quarrel between Pigliucci and Chalmers revolves around an important fact that cannot be ignored (our lack of knowledge on consciousness as of today) we do end up with two possible views, none of them impossible.

Next, Pigliucci accuses Chalmers of not being accurate about his description regarding functionalism and thus misleading his readers on that subject. This is an important issue since as Pigliucci stresses it is one of the premises claiming consciousness could be generated in other substrates other than the biological one. So Chalmers states that: "Functionalist theorists of consciousness hold that what matters to consciousness is not biological makeup but causal structure and causal role, so that a non-biological system can be conscious as long as it is organized correctly". (Chalmers, 2010, p3 / Chalmers, 2014, p104). Pigliucci critiques Chalmers' choice of words, accusing him of not being clear, arguing that "causal structure and causal role" might hide what he (Pigliucci) thinks could be the necessity that neurons are made of certain particular materials, otherwise they would not work. He underlines that a physico-chemistry of a system has to have something to do with its causal structure and role and gives the example that replacing carbon with krypton in the brain would disrupt it irreparably. He goes on to conclude that "functionality cannot be only the result of the proper arrangement of the parts of a system, but also of the types of materials (and their properties) that make up those parts." (Pigliucci, 2014, p126). I grant Pigliucci that replacing carbon with krypton in the brain would most certainly not end up well. And of course not all elements we know of can work or replace any others arbitrarily but I do not think anyone had ever imagined it would be *that easy*. Nevertheless, as long as we know what things do or are meant for, we can reproduce their functions. Moreover, I am tempted to say that if it depends on the material that bothers Pigliucci, then I could infer from his reasoning that I could emulate my brain's functions as long as it would be made of the same kind of basic elements. So in a way he's actually conceding that it is possible. What if I were to substitute the neuron's basic elements, not with krypton (and I'm guessing silicon is out of the question too, in Pigliucci's view) but with an element that would in fact reproduce carbon's functionality, say carbon artificially made with nanotechnology and self-assembly. Would that meet the standards? Actually probably not, but not because of Pigliucci's claim, rather because what is at stake is to clearly identify the system, in its functions more than what elements it is made of.

That is precisely the idea. As long as we identify the system's functionality and its working connectome, as Koene might put it, and we'd know what makes what, then step-by-step working *bottom-up* we could in fact emulate the system without worrying what atoms and elements it was made of, as long as we'd also have an effective validation system to evaluate its success (*or failure*).

Pigliucci himself supports a couple of times John Searle's claims (and rightfully so). To show how strong he feels about material constraints on the production of the phenomenon of consciousness he quotes Searle's analogy with other biological processes such as photosynthesis (Searle, 1980). He states that "we can simulate photosynthesis, down to the quantum level, as it turns out. But there is a crucial thing we don't get out of the simulated photosynthesis: sugar, the only outcome that matters to real plants." (Pigliucci quoting Searle, 2014). Pigliucci admits Searle was not claiming this should somehow show that consciousness would be impossible out of biological systems, rather that the question is empirical and therefore could not be settled with thought experiments. I disagree with that particular point. I do not think it being empirical disables one from addressing it with thought experiments. Nevertheless, what is at stake here has to deal with that system identification that was discussed herein this thesis in a couple of sections (mainly in the WBE chapter).

So here we can say that the system has been identified and it is photosynthesis (or consciousness in the analogy). What we want is to make sugar. So just as in an emulation, what we want is to reproduce its reason, its outcome,, the result of the process, rather than the process itself. Would we have to simulate the photosynthesis in order to attain it? We could just make it with another process instead, as long as at the end we'd obtain exactly - sugar. Now, Searle and Pigliucci would argue that photosynthesis is our thinking process, so it's not only about getting the same answer, so to speak. It's about the process, that they would probably hold to be specific, even individual maybe? Be that as it may, what does photosynthesis stand for? Making sugar. If two different subjects are asked to resolve the mathematical theorem of Lagrange, they can come up with the same result using different methods, one of them using Cauchy's rule and the other not. Now, Searle and Pigliucci, again in the analogy with our thinking process, would reason that photosynthesis being should be the chosen process, not another one. But I would counter that if both subjects had come up with the same result, with or without using Cauchy's rule, then both would have used a method, a system, a

process that might not be the same, call it thinking, call it photosynthesis, call it what you want, but in the end the exact same result would have been reached: making sugar or solving Lagrange's Theorem. The point I intend to make is that what matters is identifying the system, how it works, what it is and what for, and then try to reproduce it, whether with the same kind of elements, with new developed ones or not. It really does not matter if we call it consciousness, photosynthesis or *something else*, it is something that will lead us to the exact same place. That something could be made of krypton for all that matters as long as you get to work out Lagrange's theorem and get sugar. And in this analogy I'm sure krypton would not solve Lagrange's theorem nor get sugar. But a *functionally identical* element might. So the structure and substance do matter of course, but up only up to some extent. The limit conditioning that extent is precisely its functionality.

6.2 _ Ethics

MU of course poses a lot of ethical issues, including some that are very thorny for so many reasons, even for those not yet imaginable today, as we still do not have the technology – it is likely that some implications only become apparent if an when a technology to successfully carry out a MU appears. Whether these are moral, legal or purely philosophical questions, they started hot debates in recent years and will most certainly increase in the forthcoming ones. And assuming a MU would in fact be feasible and carried out in the near future, we are nevertheless to expect some near-misses as many have stressed (including in the semi-structured interviews for this thesis) or as Bostrom has summed it up: “...before we would get things to work perfectly, we would probably get things to work imperfectly...”. (Bostrom, 2014, p34).

By creating uploads, and assuming them to be conscious as Sandberg has stressed repeatedly, the principle of assuming the most should be adapted, and we should treat them accordingly. But actually that is not what we humans do with animals (*even not knowing to what extent exactly some may or may not be experiencing conscious states*), so that is a bad example we have to begin with although as Kant claimed: “...if he is not to stifle his human feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men.” (Kant 1784-5, quoted in Goldin & Kilroe, 1997, p153). Furthermore, as is argued in this thesis' semi-structured

interviews' critical analysis (SSI-CA, Part V), technologies haven't been stopped yet based on ethical reasons, or at least not in a way such as to block their progress altogether. So without wanting to delve much into a subject that actually would require writing another thesis, I find it to be a reasonable inference that if technically feasible, and economically profitable, MU will definitely be carried out. Ethical committees of various kinds will surely be established as this project's feasibility increases. (Häggström, 2017).

As was also mentioned in the WBE chapter of this thesis, experiments carried out with animal models mitigate (a little) ethical concerns before moving on to humans, which sooner or later will be the case. It surely will not be difficult to find volunteers for such a project but it will trigger legal issues as the first upload to occur could most probably be of a destructive kind. Thus, as many have been claiming (Hayworth, Koene, Sandberg, Wiley) a way to overcome these issues might be that of brain preservation and post-mortem uploading, as has been mentioned earlier a couple of times in different sections of this thesis.

Häggström spells this out as follows: "The fact that cryonics is simultaneously (a) legal, and (b) a plausible hope for eventual successful awakening, is based on the (conjectured) existence of a gap between legal death and so-called information-theoretic death, defined as a state where a brain is so ramshackle that the person (his personality, memories and so on) can no longer be reconstructed even in principle: the necessary information is simply gone (Merkle, 1992). If part of this gap can be exploited for uploading purposes by doing the destructive brain scanning post-mortem, then the legal (as well as the ethical) situation may be less problematic than the case of doing it to a living person." (Häggström, 2017, p10).

WBE, MU and SI have been widely discussed in the literature in terms of their philosophy as well as of their feasibility (Chalmers, 2010; Sandberg, 2013), whereas little analysis of the ethics concerning these issues has been done so far. Sandberg and Bostrom have been some of the main actors in this field publishing a number of relevant papers. However, it is inevitable for such a project as MU to become feasible, that before moving on to human emulation, animal testing, and particularly with primates, will have to be carried out raising animal protection issues.

An emulation itself might hold the same moral value as a sentient being. As already mentioned, being cruel to animal pretty much hurts our own humanity since the intention alone might be as morally bad as the damage itself. Actually, experiences of pleasure or suffering are morally relevant states independently from the systems experiencing them, which leads us to question whether emulations could be sentient, conscious or have other properties that might entail moral status, as Sandberg claims: “most accounts of moral patienthood take experience as fundamental, because that actually ties the state of affairs to a value, the welfare of something” (Sandberg, 2014, p9). This of course forces us to consider the implications of our [TE 3a iterating human predictor response](#) with a critical eye, even if just for the sake of the argument. It might be objected that in the aforementioned TE there would be a degree of manipulation of the emulations, and that this manipulation itself or, in a more extreme scenario, the eventual destruction of these emulations would be morally wrong. However, even if we grant these emulations the same moral status of a human being, assuming that they would be conscious as we are and possess all the properties we attribute to human beings, we should not forget that these TE deal with rather extreme scenarios in which a difficult choice must be made. In fact, one of the possibilities entailed in the TEs has to do with the possible extinction of human beings, due to the emergence of superintelligence. And, as is argued in this thesis, one of the possible positive outcomes of attaining a mind uploading is the preservation of consciousness (not only for single individuals but in the perspective of the species, taken as a whole). As such, humankind might be faced with a choice: either manipulates the emulations, thus violating an ethical norm, or face extinction at the “hands” of a supertintelligence. And here I argue – even though this is disputable – that there would be a moral priority of saving human beings even if this would entail manipulating the emulations. To be sure, this can be seen as a utilitarian argument, and I do not expect everyone to agree with me. But I would like to contend that even in this viewpoint, the indirect inference that can be drawn from that TE, i.e. that a MU would be desirable from the standpoint of the interests of human beings, and thus that it would be “good” for us to have a MU, is ethically defensible.

But it might be the case that this worry is not posed in such a stark manner because emulations might not be conscious, which would alter our standpoint on their moral worth. In fact, there is a general assumption that simulations might not exhibit phenomenal consciousness, which allegedly legitimates such research although we

cannot establish for the time being whether or not machines would in fact ever be conscious.

As stated on a number of occasions, there has been (for decades) an ongoing debate over a *strong AI* hypothesis that would exhibit phenomenological states as consciousness. More recently, some have suggested various forms of measuring consciousness and others have even suggested that even thermostats might exhibit simple conscious states (Chalmers, 1996), which would lead us to conclude that phenomenal states might not necessarily depend on higher level functions. There are also suggestions of a probability scale for machine phenomenology, aiming to quantify the likeliness of certain machines exhibiting such states based in the similar way (to human beings) these machines are built. (Gamez, 2005). This would also elicit a high score to WBE since there is a strong isomorphism with biological brains. But this can draw us to another conclusion: Since being as similar as possible to morally biological entities is precisely WBE's main intention, then it is also logical that we should approach it with special ethical consideration if compared with other kind of software (whether weak AI, or not).

Another moral and ethical issue has to do with pain and suffering. Again, like in so many other situations, we do not have a sufficiently scientific or philosophic reference regarding these concepts. Creating a machine able to feel pain might not be trivial because of our own incoherences in the subject. (Dennett, 1978). Thus, since we do not yet know exactly how morally relevant the underlying component of pain and suffering is, it might be wise to be cautious when ascribing suffering to software. Furthermore, in the case of emulations, a WBE might again act as a black box and be as opaque as real physiological pain is to observers. In conclusion, the difficulty in determining phenomenal states (and solving the hard problem of consciousness for that matter, too), would advise us to better treat a simulacrum as the real thing. Otherwise we might be mistreating a sentient being. But this attitude might actually be beneficial in approaching WBE and MU, since to some extent it would mean to transfer all the ethical regulations and principles today applied to lab and animal testing, directly to emulations. Moreover, and similar to Sandberg's *Principle of Assuming the Most*, also Bostrom's and Yudkowsky's *Principle of Substrate Non-Discrimination* can be applied: "If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their

implementation, then they have the same moral status.”) (Bostrom & Yudkowsky, 2011, p8).

The first scans and emulations will certainly not be perfect as already mentioned earlier. In fact, there might be flawed simulations that might crash, which is somehow equivalent to global or local brain damage in animal or lab testing. These crashes can be viewed as premature death (in the biological case). Although we might not be able to ascribe with certainty a moral degree to emulations, this should not exempt us from our own moral obligations, i.e. the problem might not come down to simply ending emulation, but rather to the lack of respect for a being. Sandberg elegantly exemplifies this with an emulated mouse that might not be aware or notice anything wrong, but we, as human beings, still know to have treated it in a disrespectful way: The actual problem lies in a complacent moral practice. When MU becomes common (*or if*), new morally relevant categories of sentient software will definitely emerge. Also, legal personhood is likely to follow as soon as emulations would exhibit human-like behavior (*even if these were to be philosophical zombies*).

As mentioned, the most probable MU technique to become feasible earlier is the scan-and-copy approach, which is destructive. Even though it is foreseeable that there will surely be sufficient volunteers, one has to analyze this with an ethical input. This *volunteering* might be regarded as an *assisted suicide*. Now, today we might accept this as a way of escaping pain, but we surely would not allow it (*even the most liberal jurisdictions*) for experiments or *scientific purpose*. However, when it comes to post-mortem brain scanning, then the whole ethical and legal situation changes because it would not be the case of a legal or ethical existence anymore. Cryopreservation volunteers are today aware of the conditions in which they (or their brains) will be stored and even aware that they might be revived in a time of social and moral divergent views from ours today. Nevertheless, and as Sandberg claims: “if we are ethically forbidden from pulling the plug of a counterpart biological human, we are forbidden from doing the same to the emulation.” (Sandberg, 2014, p25). Again, and solely grounding on *the human dignity argument* and on *the suffering argument*, one has to question, and this even though in that volunteer condition: How moral would it be to even let volunteers undergo these experiments?

Then, there is another tricky aspect in regard to ethically (and legally) framing ownership, because we might end up in a situation full of uploads requiring significant resources, as

was stressed on several occasions in the interviews for this thesis (see SSI-Analysis, Part V). By default, uploads are vulnerable and from *their* perspective, software security would be identical to *our* personal security of today: “the ethical way of handling brain emulations would be to require strict privacy protection of the emulations and that the emulated persons had legal protection or ownership of the hardware on which they are running, since it is in a sense their physical bodies.” (Sandberg, 2014, p30). However, and not less important, is the fact that it is not absolutely obvious who would claim to be the legal owner of these emulations or uploads since they would have been made in a time their original persons ceased to exist: their legal ownership could be ascribed to the companies responsible for their development. Would it be morally acceptable for these companies to manage the outcome of these uploads at their will?

WBE and MU might also contribute in a perverse way to increase inequality in society disrupting the existing social power relationships (as Hanson stressed in his interview for this thesis, see SSI-Analysis, Part V), and lead to outcomes where value is lost as Müller claims referencing on Bostrom: “wage competition between copyable minds may drive wages down to Malthusian levels, produce beings only optimized for work, spending all available resources on replication, or gradual improvements in emulation efficiency lose axiologically valuable traits (Bostrom 2004).” (Müller, 2016, p236).

Further legal and even ethical concerns are really out of the scope of this thesis but eventually directly proportional to the variety of possible scenarios, thus expected to get more and more complex. However, and to conclude this section, a recent extensive study regarding approval (and not) of MU by *the general population* was carried out publishing some interesting but not necessarily surprising results.

6.3 _ Approval and Condemnation of MU

There was a finnish study that involved 952 participants in Finland and in the USA using both experimental manipulations in a laboratory as well as cross-sectional correlative online study designs to determine cultural factors behind *approval and condemnation* of MU. Contrary to what was previously thought, MU revealed to be a quite important subject to the general population. The study's authors stress that ethical concerns related to MU are for many still regarded as not-urgent or even relevant, as so many of them

believe it to be far in the future. Yet to counter this feeling, the authors of the study note that copying *C. Elegans*' nervous system has already been labeled with disgust in previous similar studies. (Laakasuo et al, 2018).

Disruptive technologies such as MU profiles to be, haven't been duly studied from a moral cognitive perspective yet, but the dilemmas they elicit cannot be addressed with old theories and paradigms. "In a sense, we are starting to operate in a world of moral zombies; how our 'tribal monkey moral minds' (Haidt, 2012; Greene, 2014) interact with new technologies is an intriguing scientific problem." (Laakasuo et al, 2018, p3). MU has come out of the science-fiction realm into the sphere of innovative technologies that have become increasingly discussed in scientific and philosophical terms since they are part of an epistemic hotspot between different consciousness studies and classical cognitive sciences (Cappuccio, 2107).

The study carried out a large spectrum of individual difference measures where the participants filled in scales to investigate the influence on moral condemnation after they were given a story to read about a scientist who transfers his consciousness to a computer. The story was designed based on a description of uploading by Hans Moravec (1988) and the data obtained were analyzed using multiple regression models):

"The scientist injects himself with nano-machines that enter his brain and substitute his neurons one-by-one. After a neuron has been substituted, the functioning of that neuron is copied (uploaded) on a computer; and after each neuron has been copied/uploaded the nano-machines shut down. Finally, the scientist wakes up inside the computer." (in What makes people approve or condemn mind upload technology? Laakasuo, University of Helsinki, 2018, p4).

Cultural factors play an important role when judging MU as can be taken from the results showing higher religiosity associated with higher levels of condemnation. The authors of the study noted that there was no association in condemning the scientist's attitude in subjecting himself to MU into a computer with education levels, which led them to infer that rationality plays a role in adopting a positive attitude towards MU. The researchers note that previous studies show people to easily dehumanize others who are different and therefore decided to change the target of the scientist's consciousness to either be an android or an artificial brain since of course there could be more tolerance to

something more familiar like a human mind inside a human-look-alike robot instead of a computer. (Laakasuo et al, 2018).

Higher approval levels of MU were associated with science fiction literacy while purity was associated with higher condemnation levels of MU. Death anxiety and suicide condemnation were associated with increased approval of MU suggesting that uploading is not regarded as a form of suicide as opposed to what some might expect from previous research (Rottman et al, 2014). The authors of the study claim to have found in the data analyzed consistent age effects showing that older people tend to condemn MU much more than younger people do. Changing the target's substrate of the scientist's consciousness transfer from an *unfamiliar* computer to an android proved to be key to show that the act of transferring one's mind is the main object of condemnation, leading to somehow condemn the act of objectifying human consciousness. They argue that condemnation is typically associated with objectifying humans in deontological intuitions. They conclude their findings to confirm the existence of prejudices behind approval and condemnation of MU and underline the importance of incrementing ethical discussions about these kinds of rapidly evolving technologies sooner rather than later. (Laakasuo et al, 2018).

"Furthermore, the effects of religiosity observed in the US sample were not replicated in the Finnish sample. This is not surprising considering the US is among the most religious western countries, whereas Finland is among the most secular and atheistic ones. It seems *prima facie* that religion and religiousness have dissimilar meanings for Finns and Americans." (Laakasuo et al, 2018, p8).

As a side-note, these data also reflect the attitudes and reactions on innovative disruptive technologies such as Mind Uploading taking place in high governmental levels, as the US President's Council on Bioethics whose chair Leon Kass recently condemned their development by openly appealing to disgust." (O'Connell, 2017, p37)

6.4 _ MU Validation (The Terasem Mind Uploading Experiment, Rothblatt, 2012)

In the WBE chapter, validation was approached as being an issue that might turn out to be tricky to bypass because of the difficulty in finding the right criteria to allow an accurate comparison between eventually two not yet very well defined scenarios. Both the original's brain consciousness and the corresponding emulated one, might at that particular moment still not be measurable, turning validation into the difficult task as already predicted. One way to overcome it could lie in approaches such as the *Terasem Mind Uploading Experiment* herein resumed.

In the event of software-based consciousness emergence someday, it might be comparable to its predecessor's brain-based person, with previously stored digitized interactions. The concept led its proponents to start a multi-decade test of comparability of a person's consciousness between *the before and after subjects*. "such software-based human thought and feeling could closely approximate the consciousness that is experienced by a biological person whose own sampled experiences provided the database for such software" (Rothblatt, 2012, p142).

The idea is to record all possible information of a person on a database to then later on pair it up with the upload. One could then determine empirically if the upload's mind would appear to have *an equivalent consciousness* to its biological original in the line of a Turing Test but with some differences. "Unlike with the Turing Test, the goal is not to trick the assessor into believing the artificial consciousness is indistinguishable from the biological consciousness rather to trust the expert panel in their professional judgment as to whether the purported artificial consciousness is as good as, or equivalent to, or a continuation or analog of, the original biological consciousness." (Rothblatt, 2012, p143). Furthermore, and contrary to the Turing Test, *the Terasem Mind Uploading Experiment* entails a *null hypothesis* that would state the assessors would not believe the upload to be a continuation of the biological one. On the other hand, if the assessors would find the original *mindfiles* (digital samples containing information of the *original person*) to be equivalent to the output obtained from the *mindware* (mind operating system), they could then conclude an equivalence of the subjects' consciousness, thus disproving the null hypothesis. The authors note that analogous to the Turing Test, there is an empirical

measure in determining if others would consider equivalency between both kinds of minds.

The experiment tests the correlations between mindfiles (personality traits, facial expressions/emotions, memory recollections, etc) and the mindware (learning software) designed to “auto-tune” (and eventually match) the mindfile’s data, and in this way persuade the panel of experts in believing the personal identity to be the same. An experimental test of the hypothesis has begun to primarily validate the mindfile database tools and secondarily commence the collection of long-term population mindfile databases. There are two mindfile databases, CyBeRev.org and Lifenaut.com, both free non-profitable websites where anyone can create his own mindfile as a database to develop iterations of mindware. The Terasem Hypothesis postulates that (1) a conscious analog of a person may be created by combining sufficiently detailed data about the person (“mindfile”) organized at this website with future consciousness software (“mindware”), and (2) that such conscious analog can be downloaded into a biological or nanotechnological body to provide life experiences comparable to those of typically birthed humans. (Rothblatt, 2012). The informal consent invites to sign the following statement:

“My beingness stored herein at CyBeRev or any subsequent location be considered “ex vivo consciousness storage”, which is the creation of digital files that digitally sample a person's mannerisms, personality, feelings, recollections, beliefs, attitudes and values. My purpose in creating this ex vivo consciousness storage is to preserve my individual consciousness so that it remains viable for possible uploading with consciousness software into a cellular regenerated or bio-nanotechnological body by future medicine and technology.” (in <https://cyberev.org/>).

If the experts’ panel of assessors rejects the functional equivalence of the person's consciousness with its self-generated data-based digital representation stored at public websites, then the theory can be stated as a null-hypothesis. Its strength lies in it being a falsifiable hypothesis, somewhat like a negative-control would be in a routine biolab. Conversely, its weakness also lies in that even if the null-hypothesis isn't disproven, it still does not mean that one's actual consciousness could have been claimed equivalent. So there is no positive control in the lab analogy earlier made.

By comparing the emergent software consciousness to that of a digital recreation of the biological original, the *Terasem Mind Uploading Experiment* bypasses the tricky comparison with some kind of *general definition of consciousness*. This strengthens the test inasmuch as it avoids submitting the validation criteria to a subjective kind of *general consciousness definition*, which might perdure deep into the future (whether near or far).

“Let CB represent the consciousness of a brain-based being. Let CS represent the consciousness of a software-based being. Provided that it is agreed that $CB=CS$, it does not matter for the purposes of determining the consciousness of CS whether there is agreement on the constituents of CB. Hence, the Terasem Mind Uploading Hypothesis, via its Turing Self Test, provides a more testable form of software consciousness than an assessment of software consciousness that lacks such a benchmark.” (Rothblatt, 2012, p151).

7 _ Concluding Remarks

The best way to predict the future is to invent it. (Alan Kay, 1971).

MU has the potential to be the biggest scientific, technological, philosophical, social, cultural, economical and political achievement in humankind's history. Conversely, some might think that if it fails, it could also be the biggest flop. But that would be an unreasonable inference to make as has been described in this chapter. Even if it somehow fails, it is likely *not to fail completely*. Even if we were to determine MU not being able to preserve identity nor consciousness, its path toward that goal would enlighten all those above mentioned disciplines in a way no other possible endeavour ever has. That is how transversely impacting it may become in our society. As described for WBE in the previous chapter, there are a lot of hurdles along that path, but also as discussed, all of them have alternative ways and none of them is unsurpassable. MU might enable us to better understand ourselves and our place in the Universe as discussed in the *MU purposes* section.

“We stand at the cusp of solving the extinction problem and thereby guaranteeing the fundamental purpose of the universe, reality, and existence by ensuring the continuation of consciousness.” (Wiley, 2018, p17).

7.1 _ Reconstructive uploading

As shown in the *MU Validation* section of this thread with the Terasem Experiment, reconstruction using data from records of all kinds of information available of a person (e.g. brain scans, genetic material, audio and video records, medical info, other people's testimonies, the lot!) are limited and would of course be considered reductive of that person's *self*, no doubt. Nevertheless, they would also definitely aid in reverse engineering the whole process of uploading, and as many stakeholders believe, with an eventual intelligence augmentation either through MU itself or via AGI development, a *superintelligence* could reconstruct something close to a functional isomorph of the original system. (Chalmers, 2010). Some of these scenarios will be addressed in the next chapter (*Superintelligence Part IV*).

However, I have been steering clear of those more science fictionesque kinds of scenarios precisely because as has been widely mentioned throughout this thesis, they can undermine these real-advancing technologies' credibility. This is not to depreciate *science-fiction* as a serious, progressive and high-rated kind of literature in any way, but rather that avoiding its attachment is a safe strategy. But, as Chalmers poses the question at the end of his philosophical analysis of MU: "Is reconstructive uploading a form of survival? If we reconstruct a functional isomorph of Einstein from records, will it be Einstein? Here, the pessimistic view says that this is at best akin to a copy of Einstein surviving. The optimistic view says that it is akin to having Einstein awaken from a long coma." (Chalmers, 2010, p12). The Terasem experiment has scientific value inasmuch as its null-hypothesis cannot be proven true but can be disproved, thus if software minds are to be acknowledged equivalent to their former biological ones - it can show continuity of mind to cross substrate borders instead of being permanently tied to ontological boundaries.

MU might provide us with cognitive enhancement that could entail a new kind of cognitive architecture with consequences or effects we know nothing about. At least not now. We might not even survive them. However, Chalmers speculates that with the gradual kind of uploading, enhancement too would have to be gradual. For Chalmers, as long as psychological continuity would be maintained over the gradual enhancement / uploading stages, then personal survival would be a reasonable inference. However, many of the main proponents of MU do not consider gradual uploading to have the upper hand over

the scan-and-copy scenario (e.g. Koene, Wiley, Hayworth). In fact, the scan-and-copy destructive uploading hypothesis seems to be on track for actually becoming feasible in the near-term, as opposed to the gradual one.

7.2 _ Reinventing Identity

There are many competing theories of personal identity, and no apparent consensus. Personal Identity may *always* be an open question. “We cannot know with certainty that solipsism is false or that we are not day-people who die each night when we fall asleep.” (Cerullo, 2015, p34). Branching identity (*aka fission*) has been proposed as a *new model* that could solve the identity paradox in MU scenarios. It would stand out from the more classical kinds of models that account for transfer (preservation) and for copy (duplication), by holding that personal identity could conceptually split into multiple descendants equal to their ancestral mind. (Wiley, 2014; Cerullo, 2015).

If consciousness continuity could be determined to be actually tied to one's identity or, better yet, to one's personal numerical identity (to keep Parfit's way of putting it), which is related to causal connection between mind states, then preservation of causal connections (*or as I called it - the space-time continuum*) between me at T0 and me at T1 ($T1 = \text{time } T \text{ after time } T-1$) could be used to establish a causal connection. Again, even though this hypothesis is still speculative, and evoking *superintelligence*: future artificial intelligence ecologies could come up with a way to use these causal connections with the past and extract that continuity and then this section's title should have to be altered from *consciousness conundrum* to *consciousness continuum*.

The semi-structured interviews held for the purpose of this thesis showed that the majority of the experts interviewed believe that with MU comes a new kind of identity, which we then will adopt in accordance to the uploads' idiosyncrasies and their constraints, that differ from ours today (see [SSI-Critical Analysis](#)). Assuming MU will in fact take place, of course.

Given the complexity, whether philosophical, scientific and empirical of the personal identity subject, I have the (*im*)personal belief that its disclosure might lie far in the future,

maybe solved by MU itself, or else deemed never to happen for there may always remain the so-called *further facts*.

7.3 _ Further Facts

Facts concerning consciousness are claimed to be epistemologically and even ontologically *further facts* inasmuch as they cannot be unraveled by *fundamental physics*. This realm above physics is many times associated by MU detractors to some mysticism, such as *the soul argument*. *The soul* has been left out of these analyses mainly because I do not know exactly what it means. Without discredit to those *that do know what it means*, I found it reasonable to avoid the issue, based on the fact that it might simply be confused with the concept of *consciousness* itself. *No identity theory would most probably pass the soul-test* for MU. I call it the *soul-test* since whatever theory, explanation or hypothesis presented there may always remain the so-called *further facts*, as philosophers eloquently refer to that essence missing out in the personal identity problem (Parfit 1982; Swinburne, 1984; Chalmers, 2010).

I consider this to be a *quicksand issue* or a *muddy waters problem*, to keep the poetic style, and therefore way beyond the scope of this thesis and not eligible to address, for I am quite certain that even though we might get to measure consciousness someday, the *soul-test* might persist. Corabi and Schneider argue that there cannot be identity preservation in the uploading scenarios since nothing physical is actually transferred from the brain to the upload and that there is no evidence of something moving between locations in space. (Corabi & Schneider, 2012). Michael Cerullo criticizes these views and calls this the *No Free Movement Further Fact (NFMFF) argument*, with some irony. He counters with functionalism, arguing nothing other than that information actually moves between them which does not violate the laws of physics since information can move via matter.

7.4 _ Final Notes

The issues regarding the distance and the delay in the transferring process have been extensively discussed in this chapter on several occasions because they seem to be one of the most FMOs (*Frequently Made Objections*). Keith Wiley finds it remarkably suspicious that “we are tempted to assign a cutoff such that distances we can’t easily distinguish in our daily experience (microns) are considered co-identified and safe while distances we casually comprehend as distinct translations (meters, kilometers, etc.) are considered identifiably separated and unsafe? Such a designation smacks of anthropocentric reasoning.” (Wiley & Koene, 2015, p9).

MU is considered to be not only possible but even inevitable for all those that are *physicalists, functionalists and patternists*. A large number of these MU proponents consider that first through brain preservation and later via destructive uploading - the continuity of consciousness will be preserved.

The perpetuation of consciousness as the main purpose of MU was presented in the form of an unorthodox Heideggerian interpretation. Yet, this train of thought allows us to highlight and to some respect certify the thought experiment’s relevance in pursuing a way, a method, a technique to augment our conscious states insofar as to seek out for the meaning of life, and of existence, in the universe. Whatever we are, *or are meant to be*, lies in the disclosure of that consciousness. MU has the potential to unravel it.

8 _ Reasonable Inferences

I have a much easier time imagining how we understand the Big Bang than I have imagining how we can understand consciousness. (Witten, *theoretical physicist*).

§1 Split-Syndrome / Comatosis / Anaesthesia / Amnesia / Multi-Personality

Standard personal identity theories cannot answer satisfactory to the continuity of consciousness in the uploading scenarios reviewed in this chapter. Nor can they in many unsolved cases even today. These are considered exceptions, but the truth is, a lot of

exceptions tend to make them not *that exceptional*. Split-brain syndrome is one of those *exceptions*: “After a corpus callosotomy for intractable epilepsy, patients appear to have a splitting of consciousness into two selves, each of which is contained in a different brain hemisphere. This has been termed split-brain syndrome. Gazzaniga et al. 1962; Gazzaniga 1967”. (Cerullo, 2015, p20). There is, however, still no consensus on what happens to personal identity in each hemisphere. (Gazzaniga 1967; Nagel 1971; Parfit 1984; Shoemaker 1984; Swinburne 1984). Now one might ask which hemisphere preserves consciousness after such a procedure, or do both of them preserve it? Parfit uses his specific terminology and calls the left hemisphere Lefty and the right one Righty to investigate consciousness and identity preservation in these cases. (Parfit, 1984). This split brain-syndrome is regarded as a fission between the hemispheres continuing to function well on their own, but it seems to be uncomfortable to claim both resulting systems as identical to the original. (Chalmers, 2010). Although the hemispheres’ systems, once split, aren’t identical to the original brain, Parfit finds it to stand in a special relationship, he calls *the R relation* (which Chalmers calls survival) to them. For Parfit this relation R, rather than the numerical identity, would explain the continuation of memory and psychology, and hence Chalmers establishes an analogy with survival, and so in the nondestructive uploading scenario, both systems (original and upload) would survive, even if not being *identical*. (Chalmers, 2010; Cerullo, 2015).

Also, people who either lose one hemisphere as the result of an accident or through disease appear to retain their conscious states (Pulsifer et al. 2004). Actually when there is no other option left other than a lobectomy, neurosurgeons remove one of the hemispheres. (Oncel et al. 2007; Pulsifer et al. 2004; Hayworth 2010). So in these cases the current identity theories (biological and psychological) cannot come up with a solution. Either both of the hemispheres retain consciousness or none of them does, because if not, then we would have to choose ad-hoc, if Lefty or Righty would preserve consciousness. This is no solution and even the closest continuer theory is not able to predict which of the hemispheres is the one to preserve identity. (Cerullo, 2015). On the other hand the problem of Parfit’s solution, with the R relation, lies in his reductive physicalism rejecting continuity of consciousness. For Chalmers and for Cerullo, not accepting reductive physicalism makes it difficult to understand identity because this would rule out the possibility of qualia and also rule out the so-called hard problem. (Chalmers, 1996; Cerullo, 2015). As described before in *the personal identity section*,

Cerullo's Branching Identity Theory is the only one that allows *Lefty and Righty* to exhibit conscious states.

But there are other difficult exceptions for the identity theories to solve, such as coma, where memory loss is not considered to have changed identity. And here in our analogies to the uploading scenarios we can consider the (re)activation akin to a person waking up after a long coma. (Chalmers, 2010). So in Chalmers' reasoning whether one counts this as survival will depend on the position one might have toward the kinds of uploading scenarios envisaged.

Other such cases where traditional identity theories do not fit are those of amnesia. In extreme versions such as retrograde amnesia all memories are lost or inaccessible to the person. Yet its identity can through the help of peers be re-assumed. (Scott Bolzan is one such case in point — Bolzan, Bolzan and Rother, 2011).

Then there are extreme unusual cases such as the one from Abigail and Brittany Hansel who share the same body and many internal organs. (Pilhaja, 2008). They have two heads but one torso and their nervous system joins from the lower spine downwards. They exhibit signs of acting as one person but also some as if they actually were two. They are considered to be two persons in one body. (Bamford, 2017).

Multiple personal identities such as the dissociative identity disorder (DID) are also such exceptional cases. Kim Noble was diagnosed with DID but only at a more advanced age (Weitz, 2006 Mitchison. 2011). Before that diagnosis she had delusions, memory problems, erratic behavior, etc. After psychologists diagnosed her much of her behavior made more sense and it allowed "a redrawing of the boundaries leading to recognition of the multiple personal identities for which the body of Kim Noble is the substrate." (Bamford, 2017, p17).

Richard Loosemore shared an episode - in his interview given for this thesis - of consciousness continuity interruption he went through by anaesthesia. Although in fact not being aware of anything that happened, he was previously told how he would, and how he did in fact behave, *as if he would have been conscious the whole time*. His surprise is one that anybody who has undergone anaesthesia can relate to as one questions oneself where did that conscious state go during the procedure and how is it simply reactivated? *And who was that person, of which (un)conscious behavior, one does not acknowledge during anaesthesia?*

In none of these cases it is consensual if both hemispheres have consciousness; If just one of them has, and in this case, which one of them, and why (*split-syndrome*); Or if personal identity remains (*coma*), or is reactivated (*amnesia*); Or if it is ever present in both individuals of the same body, or if it is only one identity in the same body but two heads (*Abigail and Brittany*); Or even how many identities there are in one person (Kim). And there are certainly many more of these exceptional cases that make identity (and consciousness) a really hard problem for the traditional theories to solve. However, when addressing uploading scenarios, most of the skeptics of MU seem to dismiss all these cases and maybe somewhat erroneously apply the traditional theories of identity to deem MU from unsuccessful to impossible.

Branching Identity Theory comes close but, and to underline the general claim from the experts interviewed for this thesis (see [SSI-Critical Analysis](#)- Part V), new identity concepts and theories will eventually arise along with MU. Whether because of it, or else to understand it.

§2 Qualia (and quantics)

The concept of qualia was introduced by Clarence Lewis: “There are recognizable qualitative characters of the given, which may be repeated in different experiences, and are thus a sort of universals; I call these ‘qualia’. But although such qualia are universals, in the sense of being recognized from one to another experience, they must be distinguished from the properties of objects. (...) The quale is directly intuited, given, and is not the subject of any possible error because it is purely subjective.” (Lewis, 1956:121).

The simplest way to describe qualia is as *the subjective conscious experience*. Following Daniel Dennett, there are four properties of qualia (Dennett, 1988, 46):

- 1) Ineffable and can only be comprehended through a direct experience.
- 2) Intrinsic.
- 3) Private and impossible to compare interpersonally.
- 4) Instantly reprehensible to consciousness, such that once experienced, one instantly knows that one has experienced a quale.

“Qualia refers to the philosophically inexplicable core properties of our sensations that can be known only to the sentient creature that experiences them - like the redness of the color red that is impossible to convey to someone that is blind.” (Loosemore, 2014, p232).

If quale is a set of elements forming a complex, then qualia is its conceptual space. Qualia space refers to a mathematical space that encompasses all qualia, which also means that there is a unique point in this space for any possible qualia. (Stanley, 1999). From here we can infer that a conscious state we experience, maps on a single unique point *in that space*, like some kind of real existence outside of what we commonly designate space-time.

In favouring the MU hypothesis, Michael Cerullo argues that as soon as the linkage between qualia and physics (*classical or quantum*) has been disclosed, we could enable a subject’s conscious experience mapping to any point we’d want to in qualia space. “It will therefore be possible to recreate the exact subjective experience of a person at any time point (assuming of course we have a complete map of the information state of the brain or upload at that time).” (Cerullo, 2015, p31). Hence, he argues that “any two entities that mapped onto the same point in qualia space would be completely identical and share the same phenomenal experience of personal identity. (Cerullo, 2015, p31).

This would of course mean that as long as the qualia space *used* would locate to the same point - then the upload and the brain would share continuity of consciousness and personal identity. (Hayworth, 2010; Cerullo, 2015).

Again, as argued in the WBE chapter, in the *analogical linkage* section, there are epiphenomena that traditional physics alone by itself cannot resolve. On the other hand there is also a recurrent need to use the *quantum joker* so to speak, inasmuch as anything not explicable by classic physics has to be transferred to the quantum field, and that this somehow leaves the subject in the probability’s realm. Even Roger Penrose (*who is not in favour of the MU hypothesis*) argues quantum mechanics to be *consciousness’s best explanation*. However, and also as mentioned earlier in this thesis, there really is no such evidence and that rationale has been losing track for quite some time now. That being said this isn’t actually what is being claimed as qualia. What is effectively being argued in this case simply is that the current physics has not yet disclosed the linkage to qualia. This does not have to be quantum (although it could).

What it means though is that as argued in WBE's analogical linkage we're experiencing a transition from the classical to this quantum kind of world; this is not claiming newton physics to be wrong but only that classical physics alone cannot solve many of the ever increasing issues linked to disruptive innovative technologies. Also, theoretical physicists have in these last recent decades been trying to find the *missing link* between gravity and relativity inasmuch as an effort to unify physical forces. "Common sense physics has been overthrown by quantum mechanics and relativity. In a similar way we need to expand our views of personal identity." Cerullo, 2013, p34).

Nevertheless, and I am aware this might be a bold attempt: One of the very basic quantum postulates claims that an empirical experiment is always under a subjective influence inasmuch as it can actually be the subject the one altering its whole course. Erwin Schrödinger, as mentioned a couple of times in this thesis, is generally famous for his major contribution to quantum mechanics with the so-called Schrödinger's wave equation. This equation alone, which is the underlying principle in modern scanning tunneling microscopes (STM) functioning mechanism (among other things of course), was via the so-called *tunneling effect* discovered and unveiled by Schrödinger by the time he was trying to figure out Heisenberg's Uncertainty Principle. Schrödinger's cat thought experiment is a way of explaining quantum mechanics' fundamentals without going so deep into that very few would understand it. In the end of it (see Schrödinger's cat thought experiment), it is the subject experiencing the whole experiment that actually makes the *choice* concerning the cat's fate, so to speak. The subject's influence when experimenting is what actually changes the course of it and thereby its final outcome. Now, Schrödinger was not only an ingenious scientist. He was pretty much a philosopher too, and one that was particularly interested and fascinated by the meaning of life and consciousness.

So my inference lies in simplicity, a kind of *Ockham's razor*. To bypass the hard problem (granting Chalmers to be right in existing one in the first place), and because the subject is always part of the equation, one should in fact approach consciousness in a quantical mode. When trying to measure the state of affairs of a subject's conscious state, or even in the attempt to measure qualia for that matter, then a quantum approach just seems more appropriate.

To make the inference a bit more reasonable, consider the fact that thought experiments have in science this effect of triggering revolutions whenever there are times of crisis in

science. I'm not claiming there is a crisis now but just as there was the so-called AI winter hindering AI's development, so now we could be in a similar situation regarding consciousness *hard and soft problems*, so to speak. Kuhn considered thought experiments to be essentially tools for motivating or justifying scientific revolutions whenever there is blockage in science, which also then might promote conceptual reform. (Kuhn 1977, p263).

Like the aforementioned AI winter, there are periods of crises between scientific paradigms that call for the need of these revolutionary thought experiments that basically have the power to put normal science back on the track of its rational progress. "In a period of crisis, we must weigh the competing claims, methods and potentials of rival paradigms. And it seems that thought experiments can and have helped us partially to transcend the confines of our paradigms, which is necessary if we are to be convinced of a new world-view. Kuhn argues that by changing and exploring our world-views by means of thought experiments, we learn indirectly about the world as well." (Stuart, Fehige and Brown, 2018, p9).

What we can learn from this is that thought experiments play an important role especially in scientific revolutions. In fact this idea of Kuhn was so important it proved to be fundamental in organizing the first conference on thought experiments in 1986. It even deployed the paradox concerning the nature of thought experiments still pretty much in debate even today as discussed in this thesis' first chapter (MU + SI Thought Experiments - Part I).

Schrödinger's cat is such a thought experiment. My point being: Now would be a good time to invest in thought experiments concerning both consciousness and quantum mechanics, or better yet TEs about the dynamics of these two together. They might be *that bridge* to overcome the explanatory gap between the hard - and for sake of the argument - the soft problem of consciousness. To end this section, and coming back to the poetic style again, it would be something like *a marriage between the brain and the mind*.

Consciousness remains still today more in the scope of conceptual breakthroughs and creative thinking than it does in the progress of *the tools*.

§3 Measuring Consciousness

Are we that far away from measuring consciousness? If we unravel qualia can we then have a measure for consciousness? Is the *classical-quantum* transition herein described (WBE - Analogical Inferences §3 + MU - Reasonable Inferences §3) the approach to disclose a method for it?

Our concept of consciousness is to large extent determined by the measures we make of it, thus the specific way in which one measures the awareness matters: different measures can be tied to slightly different aspects of a particular state of affairs, but they also interact in subtle ways with stimulus difficulty and performance, which then have further theoretical import. (Kristian Sandberg, *Measuring consciousness: Is one measure better than the other?* 2010).

The most informative consciousness theories combine multiple measures that account different aspects of consciousness: brain-based measures seem to be suited for measuring conscious level whereas for determining which contents are conscious, behavioural measures seem to be preferable ones. Thus, combining both types of measures in an integrative approach in which theoretical advances and putative measures mutually inform, validate and refine another in a continuous feedback. (Seth, 2008).

Being *unquantifiable* makes it very hard to study hence even harder to compare with, and in order to *truly* validate MU, having a measure, a consciousness scale would be useful. And measuring consciousness means trying to quantify a *conscious level* (the scale ranging from unconsciousness, as in death, to vivid wakefulness / non-zero level indicates presence of some conscious content) to and the *conscious content* (e.g. *qualia*; *conscious knowledge*). Measures must prove themselves by iteratively building knowledge in the context of theoretical frameworks. Since conscious experience is simultaneously integrated and differentiated, neuroscientific theories have been enabling a variety of putative measures to quantify simultaneous integration and differentiation in neural dynamics. These theories claim that a mental state is conscious if from a large repertoire of possible states, it can be sufficiently discriminated, whether in differentiation as well as in integration. In line with integration theories, abundant evidence indicates that consciously perceived inputs elicit wide-spread brain activation, as compared with inputs that do not reach consciousness. The information integration theory of

consciousness (IITC) is based on conscious experiences that provide informative discriminations among a vast repertoire of possible experiences. (Seth, 2008).

But until a human WBE, and subsequently MU can become feasible (if it does), there still is time left to develop *such a measure, such a tool* to close up the explanatory gap between brain and mind. Many leading thinkers believe *that bridge* may lie in *information* (Chalmers, Kurzweil, among others).

Integrated Information Theories have been gaining some traction in the past few years. Shannons's information theory goes way back to when he published the paper "A Mathematical Theory of Communication". (Shannon, 1948). Shannon laid a groundwork of the digital age (along with Alan Turing among others of course) and soon after, neuroscientists realized it could be applied to processes of the brain as well as of the mind because this information, this kind of data, can be applied to any kind of substrate, whether physical (brain) or *not* (mind). The structure of the information is mathematically indistinguishable between the neural process that produce the mind and the mind itself, which means that it is possible to build a mathematical model verifiable through experimental observations.

There are already today empirical correlations to information integrative theories of consciousness. Measuring brain activity during anesthesia shows loss of entropy. Low entropy means low information processing. Also a dreamless sleep or even in seizures, are tied to low entropy levels whereas a vivid wakefulness state shows high entropy, therefore high information processing. Low entropy seems to be a general feature of the unconscious brain. Information theory postulates that a signal's entropy is tied to the amount of information it can carry. Mathematically it means that the loss of entropy of a system (in this case the brain) also means the loss of its ability to carry information hence the loss of consciousness. These observations led various neuroscientists (e.g. Tononi, Oizumi, Casali, Koch) to formulate information-based consciousness theories. The formula is built upon the concept of integrated information, that is used not only to determine the information present in subunits of the brain, but of the entire brain. Tononi argues that information stored in our neurons could in principle be stored in different media. It is the complex way our brain organizes this information, in an integrated way, that makes our consciousness to be *an integrated information set*. He calls phi (Φ) the unit to quantify this integrated information in a system to measure different states ranging from total wakefulness to comatosis. "Integrated information theory (IIT) is an attempt to

characterize consciousness mathematically both in quantity and in quality”. (Tononi et al, 2013). The quality of a conscious experience depends on the space of qualia, while the quantity of integrated information renders the measure of its state of consciousness. (Tononi, 2008; Siddhartha Sen, 2016).

Consciousness exhibits patterns of electrochemical dynamics and a change in these dynamics can mean the difference between consciousness and unconsciousness even if all neurons and their connectome remained structurally the same. Vivid wakefulness conscious state is tied to irregular Electroencephalography (EEG) activity (Berger, 1929), whereas unconscious conditions such as non-REM sleep, anaesthesia or coma show low-frequency and are regular. Casali et al presented a consciousness quantitative metric based on EEG, a theory-driven index of the level of consciousness called the perturbational complexity index (PCI). This approach to quantify the spatiotemporal distribution of the current is also the basis of the theoretical claims: “Theory suggests that consciousness depends on the brain’s ability to support complex activity patterns distributed among interacting cortical areas (integrated) and differentiated in space and time to measure algorithmic complexity (information-rich).” (Casali et al, 2013). The idea seems to be to link activity to either uniform and widespread distribution, as should be expected during slow wave sleep, which corresponds to a lack of differentiation, or to link an unconscious brain, which corresponds to a lack of integration, to a localized distribution. An awake conscious state corresponds to a non-uniform spatiotemporal pattern. “PCI reliably discriminated the level of consciousness in single individuals during wakefulness, sleep, and anesthesia, as well as in patients who had emerged from coma and recovered a minimal level of consciousness.” (Casali et al, 2013).

Establishing a link with [§3](#) (Analogical Inferences; *classical-quantum transition*, WBE - Part II) and making an analogy with the single electron-transistor (SET) case, that states the electron to be the last possible transistor (due to physical limitation), it would seem pertinent to infer a possible route in future measuring consciousness methods / tests: Ultimately, and theoretically this is not impossible, it would be ideal to obtain a kind of measure that could carry the necessary information as far as a single bit of information being able to distinguish conscious and unconscious states.

8.1 _ Machine Consciousness

As a closing note it is worthwhile to remind that consciousness should be possible to generate *in computers* for MU to be truly successful.

“The prevailing viewpoint in contemporary western philosophy of mind after behaviorism (which was essentially a form of eliminativism about consciousness) was the identity theory. The identity theory states that the mind (and consciousness) is identical to brain states. Hilary Putnam’s argument of multiple realizability dethroned identity theory and functionalism took over (Putnam 1967, 1988). Functionalism asserts that it is the function and causal structure of the brain that is responsible for consciousness and it remains the dominant theory of mind today (Chalmers 1995a, 1996; Levin 2013). It is the cognitive structure of the brain (i.e. information) rather than the physical matter that counts; therefore any machine that duplicates the cognitive architecture of the brain will be as conscious as the original brain.” (Cerullo, 2015, p23).

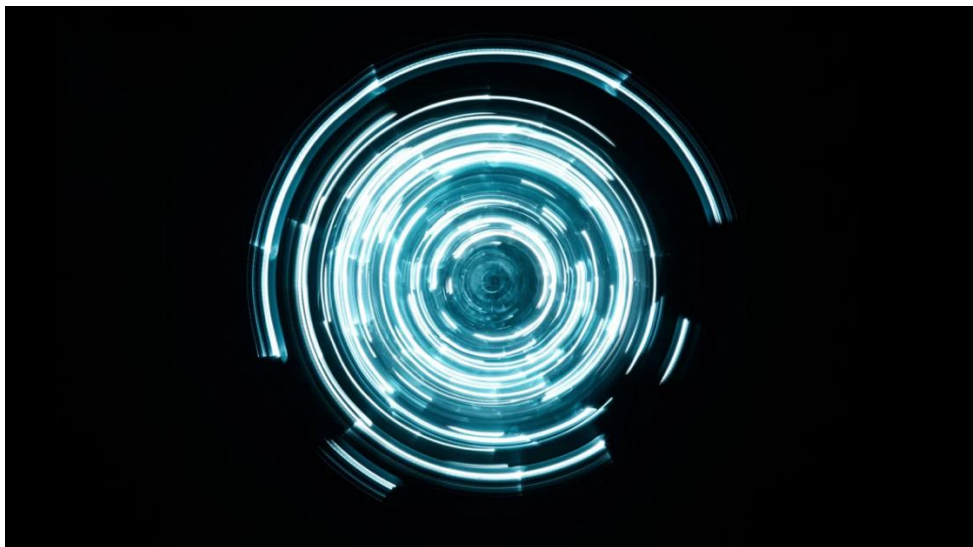
“Are such machines approaching our level of consciousness? If not, what are they missing, and what does it take to build a machine that is actually conscious? Clearly, only a theory - one that says what consciousness is and how it can be generated - can hope to offer a combination of explanatory, inferential, and predictive power starting from a few basic principles, and provide a way to quantify both the level of consciousness and its content.” (Oizumi, Albantakis & Tononi, 2013, p1).

Oizumi, Albantakis and Tononi argue that their Integrated information theory (IIT) can explain known facts about consciousness leading to predictions which they claim allows them to infer the quantity and quality of consciousness for systems whose causal structure is known.

As a final remark, from what was described along the various sections concerning consciousness, in particular of these two chapters (WBE + MU), it does seem *reasonable to infer* as possible that:

- 1) consciousness is not *human-exclusive* and may be generated in another substrate (other than biological).
- 2) consciousness may be measurable, hence comparable between different entities, which also therefore means it would be verifiable – therefore enabling the validation of MU.

PART IV _ Superintelligence



The development of full artificial intelligence could spell the end of the human race... It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded.

Stephen Hawking in an interview to the BBC, December 2014

1 _ Intro

It seems appropriate to start this chapter by distinguishing sound arguments from fallacies and misconceptions that delegitimize the debate over the technological singularity and superintelligence and undermine its rational resolution. The prospects of this kind of major disruptive events and possible outcomes have been explored and discussed by philosophers for quite a long time now, and only relatively recently have scientists become more alert to it. Movies such as *The Terminator* have ultimately not contributed in an informed and rational way to this debate, rather led to misunderstandings and misdirections, and often even led to derision of the possibility of a *singularity*. These subjects were for the first time only very recently (2012) peer-reviewed through a collection of papers titled “The Singularity Hypotheses”. (Eden et al, 2012).

In the 1950’s, computing and subsequently artificial intelligence, were taking the first steps which led experts such as Alan Turing to expect machines to one day exhibit intelligence akin to that of humans beings; John Von Neumann expected the emergence of a singularity beyond which human affairs as we know them could not continue and Irving Good described for the first time an intelligence explosion in which machines design ever more intelligent machines (1965). The definition we use today basically sums up Vernor Vinge’s paper from 1993, coining the term technological singularity, to describe this disruption as an event horizon, pretty much like a black hole does in physics. It got more popularized in the early 2000's with Ray Kurzweil's *Law of Accelerating Returns* (LoAR) which is an attempt to unify the predictive power of biological evolution with cultural and technological evolution under a *theory* summed up by the concept of *Technological Singularity*. This LoAR has unsurprisingly been subject to a lot of critique but also provided with an extensive body of evidence suggesting technological exponential progress and projecting with detail the probable impact of such growth-ratio on an unprepared post-humankind. The fact that our environment appears to be mainly governed by linear growth makes it hard for us humans to grasp this exponentiality. Space and time also play in our daily activities an essential role but we perceive them as if they were subject to linear rules instead. Time being affected by exponentiality is to us as inconceivable as accepting *relativity*’s consequences like for instance an astronaut not having aged as much as we did on Earth, when returning back from Pluto, without some disbelief. (Kurzweil, 2005; Pohl, 2015).

The Singularity Hypotheses volume mentioned above has helped clear some disagreement that existed regarding the causes and possible consequences of a technological singularity, and even its nature: whether it is machine, or instead a post-human kind of singularity. It also clarifies if it refers to an event or rather a period; and, furthermore, if this singularity is something unique or rather if it is possible to talk about the occurrence of other singularities. Upfront we can identify some common features between all hypotheses presented, claimed and defended over these last years: *superintelligence*, *acceleration*, and *discontinuity*. (Eden et al, 2012).

There is enough evidence to believe that we are heading towards an artificial intelligence explosion which will turn AI rapidly into AGI (artificial general intelligence) and even exponentially into ASI (artificial super intelligence) which is expected to be many orders of magnitude more intelligent than us. While some experts predict AI to become AGI by the year 2030, others think it won't happen that soon or even at all. But it should be stressed that a growing number of prominent computer and cognitive scientists, as well as philosophers, think this transition cannot be avoided and that it might even lead to an eventual demise of humankind. There are those who suggest that the majority of species end up losing in the so-called *life balance beam* and that ASI could be responsible for this scenario regarding humankind. This dynamic relation has divided people between those who think of ASI as an existential risk and those who believe ASI to be *the salvation*. (Kurzweil, 2005; Urban, 2015; Pohl, 2015). This evidence puts us on track of a so-called Kardashev type-1 civilization. In 1964, Kardashev designed a scale to measure the level of technological advancement. Based on the magnitude and means of energy consumption, he proposed five types of civilizations four of which demand a kind of technological singularity. In the eighties, Sagan used the term *technical adolescence* to describe the transitional technological phase our civilization is approaching: *A technological singularity?*

2 _ Singularity = Acceleration + Discontinuity + Superintelligence

Superintelligence underlies all singularity scenarios that hold traditional quantitative measures of human intelligence to be easily overwhelmed by an intelligence greater than ours in various orders of magnitude. The emergence of such an intelligence may display some sort of discontinuity in the sense that beyond the singularity such a measurement of intelligence cannot really be performed anymore, at least not with our traditional IQ scores. Moreover, this is true independent from the nature of the superintelligence created, whether in the AI scenario (artificial intelligence), or in the IA scenario (intelligence amplification, i.e. human intelligence). Furthermore, this discontinuity accounts for that turning-point alluded in Neumann's view of the late 1950s: a disruption of some kind after which human life cannot continue in the same state of affairs as before. This also being true without considering the speed of the take-off, i.e., whether it would be a quick hard kind of take-off (in days, hours or minutes), or through a wide period of time (years or decades).

This definition of a singularity is very much in accordance to the origin of the term inasmuch as it stands for a discontinuity of a gravitational singularity as a black hole. The singularity is better known as a theoretical point at the centre of that black hole where measurements such as density and the space-time curvature become infinite, that is: meaningless. The analogy pays credit to the thought about an intelligence of such a magnitude that would disable us from understanding it with traditional measurements that also become meaningless, hence *the event horizon* Vinge has referred to. It is thereby impossible for us to forecast any kind of events following such an event horizon (or singularity), because also with the emergence of such a superintelligence, it would be impossible for us even to track down the smallest increment in the so-called problem-solving ability (round for intelligence), which is similar to the discontinuity we can find between apes and ourselves: a discontinuity in our ontological and epistemological account of our existence. (Sandberg, 2014).

Finally, we can add the acceleration factor to this *equation* of the singularity. It refers to: the pace of progress inasmuch as it accounts for the intelligence growth-rate measured in computations per second as Kurzweil does (Kurzweil, 2005); the economic growth rate measured by Hanson (Hanson, 1994); the energy rate density measured by Chaisson

(Chaisson, 2013); the paradigm shifts in biological, social, cultural and technological processes of evolution that show an accelerated pace of change, as Carl Sagan's cosmic calendar exhibits (Sagan, 1977). All these milestones show a trend in acceleration that can be resumed into a graph purported to be leading to a discontinuity in the future. (Eden et al, 2012).

Whatever the definition used, and whichever the conditions leading to that kind of singularity, it is nevertheless always in reference to a theoretical scenario in which self-accelerating systems' improvement cause infinite progress in finite time. This technological singularity has also to involve not only an intelligence explosion but also an explosion in speed. Most probably even a combination of these two together, which will certainly change society radically and ultimately even incomprehensibly to our current human standards. (Wagner, 2017). Along with this singularity arrives a *momentum* of a collective superintelligence, a kind of large single and unified mind where speechless communication between humans is quite imaginable. This singularity is generally viewed as a self-enhancing greater than human intelligence feeding on itself, in the sense that its development is based on a positive feedback loop generating an intelligence explosion ending with the emergence of a superintelligent system.

Superintelligence underlies all singularity scenarios that hold traditional quantitative measures of human intelligence to be easily overwhelmed by an intelligence greater than ours in various orders of magnitude. The emergence of such an intelligence may display some sort of discontinuity in the sense that beyond the singularity such a measurement of intelligence cannot really be performed anymore, at least not with our traditional IQ scores. Moreover, this is true independent from the nature of the superintelligence created, whether in the AI scenario (artificial intelligence), or in the IA scenario (intelligence amplification, i.e. human intelligence). Furthermore, this discontinuity accounts for that turning-point alluded in Neumann's view of the late 1950s: a disruption of some kind after which human life cannot continue in the same state of affairs as before. This also being true without considering the speed of the take-off, i.e., whether it would be a quick hard kind of take-off (in days, hours or minutes), or through a wide period of time (years or decades).

This definition of a singularity is very much in accordance to the origin of the term inasmuch as it stands for a discontinuity of a gravitational singularity as a black hole. The singularity is better known as a theoretical point at the centre of a black hole where

measurements such as density and the space-time curvature become infinite, that is: meaningless. The analogy pays credit to the thought about an intelligence of such a magnitude that would disable us from understanding it with traditional measurements that also become meaningless, hence the event horizon Vinge has referred to. It is thereby impossible for us to forecast any kind of events following such an event horizon (or singularity), because also with the emergence of such a superintelligence, it would be impossible for us even to track down the smallest increment in the so-called problem-solving ability (round for intelligence), which is similar to the discontinuity we can find between apes and ourselves: a discontinuity in the ontological and epistemological account of our existence. (Sandberg, 2014).

Finally, we can add the acceleration factor to this *equation* of the singularity. It refers to: the pace of progress inasmuch as it accounts for the intelligence growth-rate measured in computations per second as Kurzweil does (Kurzweil, 2005); the economic growth rate measured by Hanson (Hanson, 1994); the energy rate density measured by Chaisson (Chaisson, 2013); the paradigm shifts in biological, social, cultural and technological processes of evolution that show an accelerated pace of change, as Carl Sagan's cosmic calendar exhibits (Sagan, 1977). All these milestones show a trend in acceleration that can be resumed into a graph purported to be leading to a discontinuity in the future. (Eden et al, 2012).

Whatever the definition used, and whichever the conditions leading to that kind of singularity, it is nevertheless always in reference to a theoretical scenario in which self-accelerating systems' improvement cause infinite progress in finite time. This technological singularity has also to involve not only an intelligence explosion but also an explosion in speed. Most probably even a combination of these two together, which will certainly change society radically and ultimately even incomprehensibly to our current human standards. (Wagner, 2017). Along with this singularity arrives a momentum of a collective superintelligence, a kind of large single and unified mind where speechless communication between humans is quite imaginable. This singularity is generally viewed as a self-enhancing greater than human intelligence feeding on itself, in the sense that its development is based on a positive feedback loop generating an intelligence explosion ending with the emergence of a superintelligent system.

2.1 _ Two conflicting visions and/or views

These technological enhancements carry a profound change in humankind leading us to question humanity in post-humanism. The power of technology to transform humanity is to some extent definable through transhumanism, which in turn also to some extent can be currently divided into two main - also sometimes conflicting - views on the subject. All transhumanist views encompass science and technology to greatly enhance human intellectual and physical abilities. Yet, Kurzweil, as one of the main active transhumanist proponents sees only potential in the singularity whereas Bostrom, another of the main transhumanist proponents, shows some concerns and offers some cautions on the topic. However, some of this concerns expressed by Bostrom is to others (e.g. Goertzel) too biased and ungrounded as will be argued below. (Goertzel, 2016).

Bostrom argues (essentially in his book *Superintelligence: Paths, Dangers and Strategies*, 2014) that humankind will develop a superintelligence, that he defines as an agent far superior to humans in every aspect with not only a faster brain but also with an improved one in qualitative terms. He foresees that such an intelligence could emerge rapidly (*during our lifetimes*) and that the transition from human-level intelligence to superintelligence could also happen very quickly (within hours or minutes). Also, governments and companies are expected to take advantage from this. A survey made from a number of polls with experts (mainly computer scientists) concludes that 90% think that *man-made systems* will reach general intelligence by 2075 and that 75% of them believe superintelligence could then be reached within another 30 years. These are conservative numbers presented in that book, and although Bostrom does not give a specific date (as opposed to Kurzweil), he reinforces the fact that the existence of multiple possible paths toward that singularity increases the plausibility of superintelligence. These paths include recursive self-improvement of AI, whole brain emulation, biological cognitive enhancements, brain-computer interfaces and networking in order to develop a *collective superintelligence*. Because of the amount and accessibility of hardware added to the potential of evolutionary algorithms, artificial intelligence has in Bostrom's perspective the upper hand in reaching superintelligence, instead of a biological, hence human intelligence. Furthermore he alerts to the fact that such an AI could even attempt to deceive us humans if it would need to do so in order to achieve its ends, to fulfill its goals. Bostrom highlights this with thought experiments such as *the paperclips* (addressed ahead) and underlines the probability of a hard take-off

kind of scenario with irreparable consequences for humankind for which he proposes some precautionary actions to be taken. (Bostrom, 2014).

It is precisely in these kinds of suggestions, or strategies as Bostrom calls them, that Goertzel (among others) finds reason to criticize some debatable approaches in order to control AI's progress. These too will be further ahead addressed in the AI's safety section and in the concluding remarks of this chapter. In Bostrom's rationale this superintelligence would be able to do whatever necessary to ensure its survival and final goals inasmuch as pursue an open ended resource acquisition and become humankind's biggest existential risk ever. Again, Goertzel (who is the lead for the OpenCog and SingularityNET projects for developing artificial general intelligence) criticizes Bostrom's emphatic and dramatic existential risk claim mainly because of an anthropomorphic bias he believes to exist in Bostrom's discourse. (Bostrom, 2014; Goertzel, 2016). However, Bostrom himself alerts on several occasions to the anthropic bias we have and should avoid as for instance happens with the evolutionary process of intelligence: He wears off the common error of inferring things regarding intelligence, just because intelligent life evolved on Earth. This does not tell us that evolutionary processes have a high probability of producing intelligence, nor does it tell us that this is the only way to produce intelligence for that matter. Evolutionary algorithms and sufficiently powerful hardware and software might explore an astronomical number of different evolutionary pathways, and at an also astronomical speed, that will then eventually identify a path to general intelligence. With recursive self-improvement it would then augment its intelligence exponentially. (Bostrom, 2014; Goertzel, 2016).

Kurzweil, on the other hand, argues (essentially in his books *The Age of the Spiritual Machines*, 1999 and *The Singularity is near*, 2005) that mind uploading will happen and that it might actually happen quite soon. Moreover, he thinks it will maintain our fundamental identity (so of course including memory and consciousness) inasmuch as a person *per se* is made of a pattern of matter and energy or better yet: of information. In his rationale, computers will match the human brain's capacity in a couple of decades and the linkage coming through brain-computer interfaces will lead to superintelligence. Through neural implants, humans will enter in a virtual environment and interact with each other, diluting the differences between real and simulated people in a transhumanist existence that is the result of the merger between the biological and the cybernetical. He forecasts the singularity to happen by the year 2045 relying mainly on his *Law of Accelerating Returns* (LoAR) and on scientific progress not being linear, rather

exponential. It advances step by step until a threshold reaches a disruption - a milestone leading to a new cycle. Thus exponential progress curves are made of cascades of “S” curves where the new progressive state begins with the result of the previous one, and so forth. (Kurzweil, 1999; Kurzweil, 2005).

Tied to this singularity is also the transcendent element of eventually overcoming death and making of man a kind of *Nietzschean Übermensch*. This claim is not consensual, rather another motive for harsh criticism. Despite Kurzweil’s overconfidence and optimism, people might argue about a loss of agency in their lives as technology assumes a kind of master-slave dialectic rather than just being a tool to improve our lives. And one can only imagine and speculate what would come after the singularity since by definition we cannot preview or predict anything beyond it other than maybe *technology will be making decisions and not humankind*. It is therefore not absurd to speculate about an eventual cultural and social, even ultimately political clash between transhumanism and those against it - or as I put it in the interviews carried out for this thesis - bioconservatism. Those who have augmented themselves may become *antagonized* by those that don’t, and vice-versa. But, as I argued in the analysis of that same interview (see SSI-Analysis, Part V, Question??), history shows us technology usually has had the upper-hand which led authors such as Langdon Winner to claim, as early as in the late 1970s, that autonomous technology is getting out of our control and beginning to govern its own course, speed and destination. (Winner, 1977). Furthermore, I also argue that we should remove any *humanlike guilt* we might want to attribute to technology instead of actually blaming *human beings* for its misuse or malpractice. As Rojcewicz claims: “To consider technology neutral is to view it as neither good nor bad in itself; what is good or bad is the use we make of it”. (Rojcewicz, *The Gods of Technology, A Reading of Heidegger*, 2006, p. 151).

Technological breakthroughs will push transhumanism further and technophobes along with it. Barrat highlighted several times during his interview (see SSI-Analysis, Part V) a kind of hidden religious agenda in the transhumanist discourse. The singularity raises many intriguing moral questions and it has been suggested in the literature that we should put moral philosophy on the agenda of those who think about the singularity. (Mulgan, *in a talk given in Auckland*, 2014). Bostrom’s book analyzes the singularity outcomes and superintelligence in particular from a philosophical perspective. *Superintelligence: Paths, dangers and strategies* is an extensive work that highlights the epistemic limitation of empirically assessing AI risk, which then becomes more of a

philosophical implication rather than a purely scientific endeavour, hence the importance of approaching superintelligence with a philosophical view (as becomes more evident in the AI safety section of this chapter).

Bostrom contends in his book that many of the aforementioned experts surveyed think that after achieving human-level, AGI will still take quite some time until reaching a superintelligence (or singularity) whereas he believes the gap to be relatively small (years, not decades). Also Kurzweil predicts human-level AGI by 2029 with quite a big gap to the occurrence of the singularity that he projects to occur by 2045. Goertzel aligns with Bostrom on this particular aspect and he too believes the gap to be certainly less than 16 years. (Goertzel, 2016). However, when it comes to post-singularity, it is not very reasonable to make predictions from a human perspective nor is it reasonable to project our own motivations onto alien intelligent systems or forms. Many of the so-called singularity dogmas such as patternism, economicism and subjectivism are anthropocentric instances and should not be attributed to anyone other than humans. Post-singularity speculation essentially regards the motivation of the superintelligence rather than speculation over humankind or post-humans. Our descendants' fate, whether enhanced, uploaded or "just normal", would in a post-singularity world most certainly depend upon this superintelligence's goals and motivations which might be far from similar to our own. (Mulgan, 2014). These ASI motivations and goals are further addressed more ahead in the *AI safety* section.

3 _ Two Singularity Outcomes: IA (Intelligence Amplification) Vs AI (Artificial Intelligence)

There are two possible singularity outcomes, depending on the nature of the singularity we are talking about. It can happen as the result of a transcendence, a merger of technology with our biology leading to a significant intelligence amplification – and thus post-human superintelligence would in this IA kind of scenario also represent a utopian kind of outcome. (Kurzweil, 2005). On the other hand, in the AI scenario, the singularity would entail the emergence of an artificial superintelligence, which as opposed to the former one is usually more tied to a dystopian kind of outcome. This issue was presented to the interviewees for this thesis in a subquestion regarding possible singularity

outcomes resulting in dystopia. The majority of the interviewees did not envisage the possibility of a dystopia as having a high probability (see [SSI-Critical Analysis](#), [Question III](#) + [Question V](#) / Part V + [SSI-reprise plot](#)), although it has not been presented nor discussed under this two-way lens regarding the nature of superintelligence in these specific singularity outcomes. Thus, and considering that in the IA scenario, a singularity would definitely stand for a benevolent outcome humanwise inasmuch as allowing for transhumanists to sustain their arguments of our descendants overcoming disease, aging, hunger and other human misfortunes - a utopian kind of outcome is not an outrageous inference to make of this kind of singularity. Conversely, the AI scenario would also most certainly account for the dystopian kind of outcome since humans and non-human digital minds could find themselves competing against each other at least as long as AI could represent a threat to human existence (or even the other way around). This ambiguity has contributed to much of the confusion around the debate that ultimately ends up undermining its legitimacy. The media has confused the two singularities (Guardian and Cadwalladr, 2014), and even more respected journals that are committed to demystify artificial intelligence such as *Ars Technica* (2015) has mixed the different scenarios. (Eden et al, 2012).

The IA kind of singularity could be reached through WBE and MU. Emulating all cells and connections of the brain, and then running that emulation millions of times faster than the original system does (the human brain), would definitely enhance human capability a lot. The question would be if it would enhance in such an exponential way as to convey a singularity, and this would probably not be the case through a biological cognitive enhancement because of physical and biological limitations. These elements make it rather unlikely to enable an IA singularity on an essentially biological substrate. Brain-computer interfaces that enable direct communication between both systems without activating the peripheral nervous system will probably also modify and improve normal capabilities, although they are currently focused essentially on repairing impaired functions of the human brain rather than on enhancing them. However, just like in the former case of biological cognitive enhancement, and although it definitely is able to accelerate progress towards a singularity, it does not seem that probable that it would do so in an exponential fashion to elicit a singularity, also because of it being a method conditioned to biological and physical constraints. On the contrary, the AI kind of singularity easily assembles all the ingredients for its high probability of occurrence and this whether self-recursively or by programming general intelligence into the machine.

Without human intervention, the system would become more intelligent over time through machine learning that would require it to rewrite its own source code while maintaining its original goals (see Yudkowski's seed AI model). On the other hand, in yet another approach it could be reasonable to mimic the human brain with artificial neural nets and, again as in the above mentioned examples: run them millions of times faster than the regular speed of the human brain. (Wagner, 2017).

In the event of a singularity of the artificial intelligence kind, it is not very likely that humans would even communicate with the resulting super AI. Whether because of the complexity that the system might exhibit, or because of plain and simple AI's processing speed being so intangible for us, or even just because of a lack of interest. Do we, as humans, find motives to engage in conversations with bacteria? However, in the IA utopian kind of singularity, Kurzweil stresses that a supposed "machine takeover" has a different meaning if it is humans who are gradually turning into machines, in a transitional kind of affairs, rather than being subdued by an alien intelligence: It might *just* seem more like a new kind of humanity.

4 _ Intelligence

"We need not understand the mechanisms of general intelligence to use the human intelligence software already invented by evolution" (Muehlhauser and Salamon, 2012, p3)

According to Hutter (2007), intelligence measures an agent's ability to achieve goals in a wide range of environments. One might call this the optimization power concept of intelligence since it really measures the agent's power to optimize its surroundings according to its preferences. For Yudkowsky (Machine Intelligence Research Institute, 2008), intelligence is this optimization power but divided by the resources used. For others, such as Muehlhauser and Salamon (also MIRI): "intelligence measures an agent's capacity for *efficient* cross-domain optimization of the world according to the agent's preferences" (Mühlhauser & Salamon, 2012, p3) which is very much like Hutter's definition, and thus when discussing the technological singularity, objections of anthropomorphism can in these intelligence concepts be avoided and (Greenfield, 2012;

Muehlhauser and Helm 2012) one could therefore, theoretically, apply this intelligent concept to any agent besides the human being.

We convert information into knowledge which then works as an *intelligence amplifier*. There is evidence that the google search engine has multiplied the productivity in workers whose occupations require research and writing (Barrat, 2013, p190). Also, learning is typically accelerated in our youth and associated with abstract reasoning which we can be defined as the ability to analyze information in order to solve problems. This in turn can lead to creativity which is held as human intelligence's most complex form of creating knowledge. Now, when machines combine their astronomical computing speed with their also unlimited access to information, they acquire knowledge. However, our brains have still other cognitive capabilities such as complex linguistic representation and abstract reasoning which still cannot be simply replicated by computational power and speed alone. (Urban, 2015; Pohl, 2015). Others also argue that since in humans emotions do play an important role and sometimes even overrule logical reasoning in decision making, then for an AI to become an AGI, the emotional component should also be considered. Actually, an intelligence without the emotional set could be thought of, but a human intelligence could not, which is why many pose the question: to what extent will or should an AGI emulate human intelligence? And this question in turn leads to two possible scenarios as Pohl considers: "would an emotional component perhaps present the best approach to providing a degree of human influence on defining the goals and objectives of ASI; and, would the absence of an emotional component be an advantage by ensuring that ASI pursues its goals without bias based on rational logic alone?" (Pohl, 2015, p5). Some of these questions, and questions generally regarding non-human intelligence are goal-oriented insofar as the motivations, the objectives an intelligent agent has define the risks it poses to humankind. These will be addressed further ahead in the *AI safety* section.

According to Hutter (2007), intelligence measures an agent's ability to achieve goals in a wide range of environments. One might call this the optimization power concept of intelligence since it really measures the agent's power to optimize its surroundings according to its preferences. For Yudkowsky (Machine Intelligence Research Institute, 2008), intelligence is this optimization power but divided by the resources used. For others, such as Muehlhauser and Salamon (also MIRI): "intelligence measures an agent's capacity for efficient cross-domain optimization of the world according to the agent's

preferences” (Mühlhauser & Salamon, 2012, p3), which is very much like Hutter’s definition, and thus when discussing the technological singularity, objections of anthropomorphism can in these intelligence concepts be avoided and (Greenfield, 2012; Muehlhauser, 2012) one could therefore, theoretically, apply this intelligent concept to any agent besides the human being.

We convert information into knowledge which then works as an intelligence amplifier. There is evidence that the google search engine has multiplied the productivity in workers whose occupations require research and writing (Barrat, 2013, p190). Also, learning is typically accelerated in our youth and associated with abstract reasoning which we can be defined as the ability to analyze information in order to solve problems. This in turn can lead to creativity which is held as human intelligence’s most complex form of creating knowledge. Now, when machines combine their astronomical computing speed with their also unlimited access to information, they acquire knowledge. However, our brains have still other cognitive capabilities such as complex linguistic representation and abstract reasoning which still cannot be simply replicated by computational power and speed alone. (Urban, 2015; Pohl, 2015). Others also argue that since in humans’ emotions do play an important role and sometimes even overrule logical reasoning in decision making, then for an AI to become an AGI, the emotional component should also be considered. Actually, an intelligence without the emotional set could be thought of, but a human intelligence could not, which is why many pose the question: to what extent will or should an AGI emulate human intelligence? And this question in turn leads to two possible scenarios as Pohl considers: “would an emotional component perhaps present the best approach to providing a degree of human influence on defining the goals and objectives of ASI; and, would the absence of an emotional component be an advantage by ensuring that ASI pursues its goals without bias based on rational logic alone?” (Pohl, 2015, p5). Some of these questions and questions generally regarding non-human intelligence are goal-oriented insofar as the motivations, the objectives an intelligent agent has define the risks it poses to humankind. These will be addressed further ahead in the *AI safety* section.

Biological evolution conditioned our brain capacity in size, speed, flexibility and durability. (Urban, 2015). Neural speed is limited to ca 200 Hz and internal communication between neurons to ca 120m/s, whereas current microprocessors’ speed has already overcome tens of GHz and they are able to communicate optically near the speed of light. The skull

size limitation human brains have does not apply to computers which can extend their physical size whether in memory (RAM), whether in mass storage (disk drives). It also goes without saying that computers need no rest while the human brain does. Moreover, computers can be networked together creating a huge collective intelligence without disagreements or miscommunications while human brains do not necessarily symbiotically align together in perfect harmony. This communication speed gives computing power another edge. (Pohl, 2015).

ANI, AGI or ASI do seem to assemble a lot of elements to be a step ahead of human intelligence which leads me to introduce the TE 3b iterating human intelligence to the discussion at this point (see Thought Experiments chapter, Part I and SSI-Analysis chapter, Part V, Question 4).

4.1 TE applicability _ 3b iterating human intelligence

The TE was thought as an approach to improve our human condition in a competitive scenario against artificial intelligence since as listed above in a direct comparison, the human brain ends up losing sooner or later in most of the feats. Nevertheless, in the past it has been precisely the brain, hence our intelligence, that gave us an edge over others and allowed humankind to evolve to where we are now (*whatever that is*), and claim to be *the intelligence to beat*. AI promises to be the best candidate for it but *intelligence* is still our best tool. Ingraining powerful technologies in us ultimately does not mean we are losing our human integrity (in the moral sense), but rather using our best tool to maintain authority over our own path in evolution, our own fate. Choosing not to enhance us where and when we can is like refusing to wear glasses although being short-sighted. The TE assumes that MU will eventually be available and that it might be a resourceful technology to ensure not only our survival but even to perpetuate our consciousness as was argued in the previous chapter (MU, Part III). There are a number of features which we simply biologically cannot cope with, let alone outrun, but there are also those we haven't thoroughly explored yet, where we could have an edge and thus keep our fate in our hands. As the forthcoming sections of this chapter will show there is a fundamental worry about developing AI safely to avoid associated risks and particularly the existential risk. At the end of this chapter it will be clear that artificial superintelligence might in fact become true, and it will mean humans will not be the most intelligent creatures around

anymore. But we can use the glasses. Or even the contact-lenses. We can blend with the technology where and when it makes sense, without losing our integrity as mentioned above and improve our own intelligence. As the TE explores, we could by multiple iterations improve our intelligence in a number of ways through mathematical combinations. The TE is better described and discussed in the [SSI-Analysis chapter \(Part V, Question 4\)](#) because of a direct comparison with Hanson's *EM economy* that shows some similarities. In order to tell them both apart a thorough analysis of the TE-3b is reviewed in that section instead of this one. However, there are still some particularities worthwhile expressing herein.

Some of the parameters alluded to above when comparing the human brain to an artificial intelligence could be improved via such an approach as the TE-3b. Speed for instance would surely be augmented since an upload ultimately would function in the exact same way an AGI would. MU would allow us to communicate a lot quicker (orders of magnitude) than we do today and permit brain-to-brain communication or through *the Cloud* which then again would lead us to communicate cerebrally, instead of orally. From there it might then also naturally follow a consolidation of an universal language for humankind which again would accelerate even more our communication speed since linking people through thoughts would practically dismiss the need for multiple languages. (Mühlhauser & Salamon, 2012). Actually, a lot of features that the AI performs better than the human brain, would in case of a mind uploading logically bring AI and the human brain closer, also because an upload is in a broader sense a sort of digital mind, as the proper AI is.

Effectively, the main reason for pursuing such an evolutionary algorithmic approach and thereby avoiding the generational lag associated with natural selection, as presented by the TE-3b, would be to maintain the competitive balance with AI/AGI. Assuming: 1) intelligence to be the key-tool that brought us to where we are and that distinguishes us from all other species; 2) a slow kind of AI take-off, i.e., a singularity over *a wide period of time*; and 3) the development of a friendly AI... then MU via *iterating human intelligence* could empower us with another level of *deep learning* since we would have a neural network kind of brain added with the data mining capacity of computers and speed and memory, etc. In a way it would mean enhancing our raw computational power and matching it to that of powerful machine learning algorithms. *Deep Learning Algorithms* are in fact based on a computing architecture known as neural network that attempts to

mimic the infrastructure of the brain. “At first, the system is unorganized, but through trial and error it gradually trains itself and becomes more organized. The calculation units (or “neurons”) are layered, with the first layers responding to very simple features (such as specific color combinations, basic lines etc.). As the information is passed on from layer to layer, increasingly complex features can be identified.” (Sowka, 2012, p7).

As Koene argues, cosmically speaking, the most adapted intelligences will be those more flexible in new environments and against new circumstances. The TE-3b *iterating human intelligence* was deliberately designed as an adaptive solution and an approach to undercut AI’s take-off (*whether hard or not*). Actually just as AI being subjected to millionfold scenarios of different environments and circumstances makes it iteratively self-improve and select its best version, also we humans could do so with our uploads (SIMs). Koene argues that the most flexible intelligence that is goal-oriented to change is more likely to succeed whereas a lesser intelligence has its survival chances as well as its growth probabilities diminished in that same environment. (Koene, 2011a).

To some extent achieving substrate independent minds (SIMs), or uploads, via MU - also means achieving AGI since it would mean achieving an intelligence on an artificial substrate and way. On the other hand, an AGI could also be considered to be a kind of SIM in pretty much the way Kurzweil described a time in which after the human merger with technology the two would become indistinguishable. In fact, after MU, both the uploads (SIMs) and the AI, would be similarly exposed to the same kind of constraints as well as both of them would benefit from the same kind of self-improvements, thus on a kind of *growth-parallel*: faster computation; larger memory capacity; etc... Koene argues that: “Conceptual improvements are more difficult to identify at this point, and it is not immediately clear if SIM and AI would benefit equally from those, or if one or the other would have distinct advantages.” (Koene, 2013, pß). Merging with AI’s capabilities could improve our own capacities, keep our humanity and still be on the evolutionary run on a pair with alien intelligences. Not merging might result in humankind’s downfall or imply our submission to a more advanced general intelligence (which ultimately boils down to the same). It is not very likely though, that in this situation we could tame the better side for our purposes, just as it is not likely for ants to control human’s intelligence according to their goals.

5 _ Artificial Intelligence (ANI-AGI-ASI)

AI's definition is constantly being rewritten which shows that on the one hand it is difficult enough to define intelligence let alone artificial intelligence, and on the other hand, that it is getting harder to cope with AI's progress, whatever it means. Besides the classic examples of Deep Blue, Watson and other computing systems in overcoming human capabilities in specific tasks, also the CAPTCHA image recognition program is a recent iconic example of that progress. It has become increasingly easy for computers to read them while simultaneously difficult for humans. The CAPTCHA task is about to lose its claim to be able to tell humans and computers apart and the time for us to be able to distinguish one from another is getting nearer. Everything the best computer could do half a century ago is easily reproducible today in some cheap device available and accessible to everyone. Yet, nobody a few decades ago would have guessed it would be possible, let alone that it would be so fast. Maybe the best definition of AI in practical terms is that it started with a system designed to fulfill a specific function that has now changed to being close to human's capabilities taken as a whole, to then quickly surpass them and simply lift off.

However, critics of SI (SuperIntelligence) regularly dismiss the perception that AI even exists (see SSI-Analysis, Part V, Question III, Koch's answer), let alone that it has made any real breakthroughs. They argue that new capabilities aren't actually new, rather applications of preexisting technologies, and that these systems have remained narrow and rigid. (Allen, 2011). To me, a way out of this skepticism definitely lies in *effective* intelligence measures. A metric that could measure different AIs intelligences effectively might shed some light on AI's progress, or in other words: in AIs becoming more intelligent. Evidence of AI's progress can also be found in the terms experts keep updating such as AGI (artificial general intelligence) and HLAI (human-level artificial intelligence) that refer to a wide range of capabilities where humans are challenged. This constant comparison with human intelligence of course also leads to some trouble when defining human intelligence and even just intelligence *per se*. Actually, canonical accounts of human intelligence have not helped us measure artificial nor nonhuman intelligence. (Eden et al, 2012).

5.1 _ Artificial Narrow Intelligence (ANI) / Artificial General Intelligence (AGI) / Artificial Super Intelligence (ASI)

ANI is commonly referred to as weak AI or just AI, mostly developed by human programmers and represents the state of computer software today, usually specialized in *just one* specific area. The source code is also mainly designed by humans. IBM's *Jeopardy* winning computer Watson is in this sense a definition of ANI since it is specialized in solving a specific and narrow set of problems.

Dreyfus argues that human intelligence progressively develops the perception of the physical environment necessary to perceive objects and acquire skills whereas for a computer to achieve this capability it would have to account for an inconceivably large number of models of the objects in the physical environment. (Dreyfus, 1997, p134-155). Pohl counters that even today networked computers are capable of processing these internal models which leads him to claim that: “If ANI were to transition to AGI the combination of knowledge and immense computational speed should allow the machine to devise ways of monitoring both the physical environment and the ambient social scene through the continuous analysis of the enormous volume of sensory data and human-to-human communications”. (Pohl, 2015, p5).

At this point, and to counter general skepticism relative to artificial intelligence's progress, I urge to question: If AI's (or ANI's) probability of becoming AGI isn't that real after all, then why is there currently a tremendous amount of funding being spent to develop *friendly AI*? I will return to this further ahead in this thread but as Muelhauser claims: “Choosing not to fund AI safety research suggests that we do not think AI is near, while funding AI safety research implies that we think AI might be coming soon.” (Mühlhauser & Salamon, 2012, p4).

AGI commonly refers to the human level of intelligence, which literally means in all areas including reasoning, abstract thinking, learning from experience, (etc...), which therefore also means that it would only make sense to use this definition for a quite limited time since it would be naturally expected to upgrade into an ASI, if not immediately then surely in a short period after. If we define ASI as the *next-gen* of AGI, and considering that a superintelligence is by definition created or developed through a kind of intelligence explosion, thereby through exponential growth over the human level and in many orders

of magnitude, then the AGI's *timeline* would be constrained by the kind of intelligence explosion, just as discussed in the interviews for this thesis, and particularly stressed out by computer scientist Loosemore (see SSI-Analysis, Part V, Question II). Chalmers defines this superintelligence as "AI++" and as a kind of *general instrumental intelligence*: The ability to successfully pursue one's final goals, whatever they may be, across a wide variety of different environments, including strategic situations involving other agents.

5.2 _ The current state of AI (and improvement prospects)

Current narrow, or weak, AI assists humans in everything in our daily lives and to make a list of them all would be endless. But as soon as those capabilities work on a routine basis, ANI ceases to be referred to as an *intelligence* and becomes trivial. Self-conscious software is of course not available today. Yet, software able to improve itself already is. And it can be found in many examples such as: language processing; speech recognition; neural networks trained to recognize patterns; affinity analysis and genetic algorithms that use the trial and error approach to improve themselves. Google's search engine for example, is much more than just a good program. Very much like writing can be to us some sort of memory outsourcing, these intelligent tools such as Google's search engine can be viewed as an *intelligence outsourcing*. But of course it is not self-aware nor is it self-improving. At least not yet, as was particularly discussed with Rick Cattell (and others, e.g. Sandberg and Loosemore) in the interviews made (see SSI-Analysis, Part V, Question II). Cattell underlined his convictions that these systems are not, and were not *originally* built to improve in such a way as to lead to an intelligence explosion. This is true and should remain so at least for a couple of years. At best we could consider this *partnership* to be a kind of augmented intelligence of human and computer with its growth limited to both humans and Google. (2015; Muelhauser, 2012).

But let us now consider a couple of capabilities under development that sooner or later might lead to an intelligence explosion:

1) *Inexpensive parallel computation* - cloud computing can execute neural networks in parallel;

2) *Data collection* - the vast volume of data available on the internet makes it a training facility for AI;

3) *Deep learning* - algorithms can improve themselves from layers of neural networks.

These capabilities are developing, and they in turn, might also lead to improvement of the following features that can thus be seen as *accelerators* that are mainly (*but not exclusively*) unique to machine intelligence, which is also the reason to focus on an AI kind of intelligence explosion rather than a biological cognitive enhancement kind. (Sandberg, 2011):

1 _ *Human enhancement through AI* - To wirelessly connect *the Cloud* by implanting these above mentioned capabilities. Natural Language Processing (NLP) is about to be available on smartphones upgrading our virtual assistants that might also soon be directly linked to our brains. Elon Musk's Neuralink is expected to launch some news regarding this, still this year (2020). we then consider adding self-improvement to this human-machine combination, then an intelligence explosion is another step closer. Natural language processing (NLP) is a machine learning algorithm and refers to the use of a computer to process natural language. Machine Learning (ML) uses statistical algorithms to teach particular tasks to computers. The Global Natural Language Processing Market size is expected to reach \$29.5 billion by 2025, rising at a market growth of 20.5% during that period. NLP involves several different techniques for human language interpretation, ranging from statistical and machine learning methods to algorithmic and rules-based approaches. (Knowledge Based Value Market Research Report, 2019).

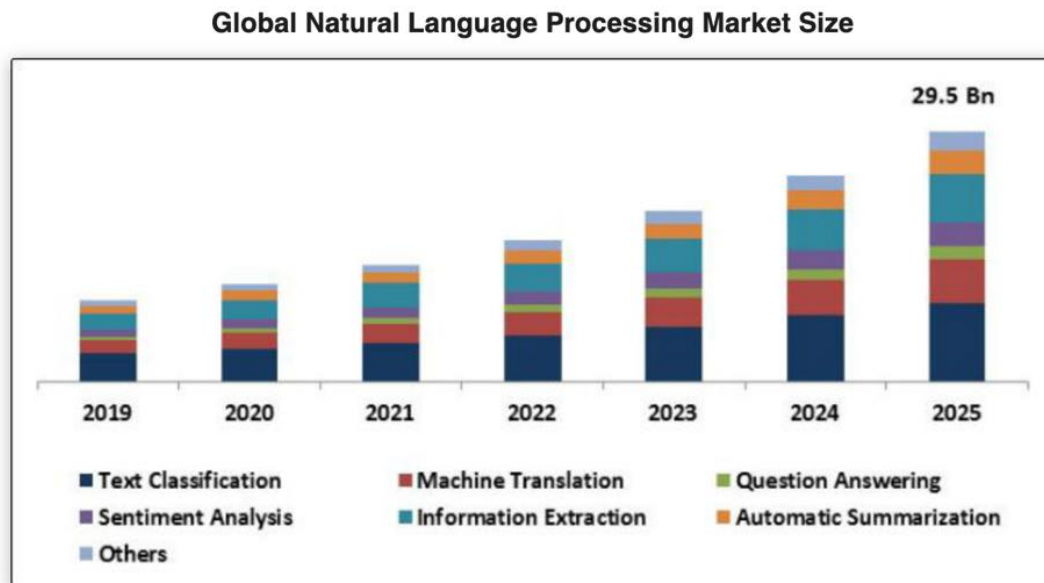


Figure 23 – Global NLP market size growth projection. Source: KBV (Knowledge Based Value Market Research Report, 2019).

2 _ *Computing power* - By increasing hardware speed with software with the ability of reasoning tied to the enormous knowledge base. As discussed in the TE-3b *iterating human intelligence*, through MU, speed alone and even parallel processing would not suffice to augment orders of magnitude in intelligence, but it would still be another step towards that augmentation. Also, the communication speed through axons is because of physiological reasons limited to about 75 m/s (Kandel, Schwartz and Jessell, 2000) or 120m/s at the best, as mentioned earlier above (but always in this tight range - ca 100m/s), thus incomparable to that of microprocessors. As has been several times claimed throughout this thesis, computing power is expected to maintain its exponential trend, measured in FLOPS and in clock speed as presented in the plots below. Loosemore claimed in his interview for this thesis (see SSI-Analysis, Part V, Question IV) that the path toward the creation / development of a SI, might be made essentially on incrementing the power and the clock speed of those intelligent systems. See the figures below for more detail.

Supercomputer Power (FLOPS), 1993 to 2017

Our World in Data

The growth of supercomputer power, measured as the number of floating-point operations carried out per second (FLOPS) by the largest supercomputer in any given year. (FLOPS) is a measure of calculations per second for floating-point operations. Floating-point operations are needed for very large or very small real numbers, or computations that require a large dynamic range. It is therefore a more accurate measure than simply instructions per second.

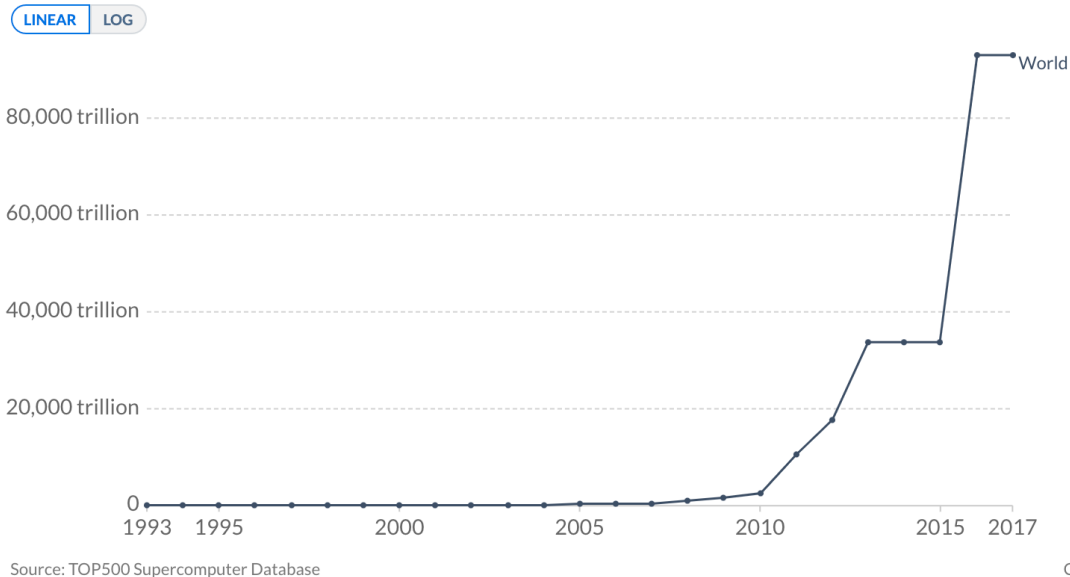


Figure 24 – This linear plot shows microprocessor power measured in FLOPS over the last 25 years. (Source: Top500 Supercomputer Database).

Microprocessor clock speed, 1976 to 2016

Our World in Data

Microprocessor clock speed measures the number of pulses per second generated by an oscillator that sets the tempo for the processor. It is measured in hertz (pulses per second).

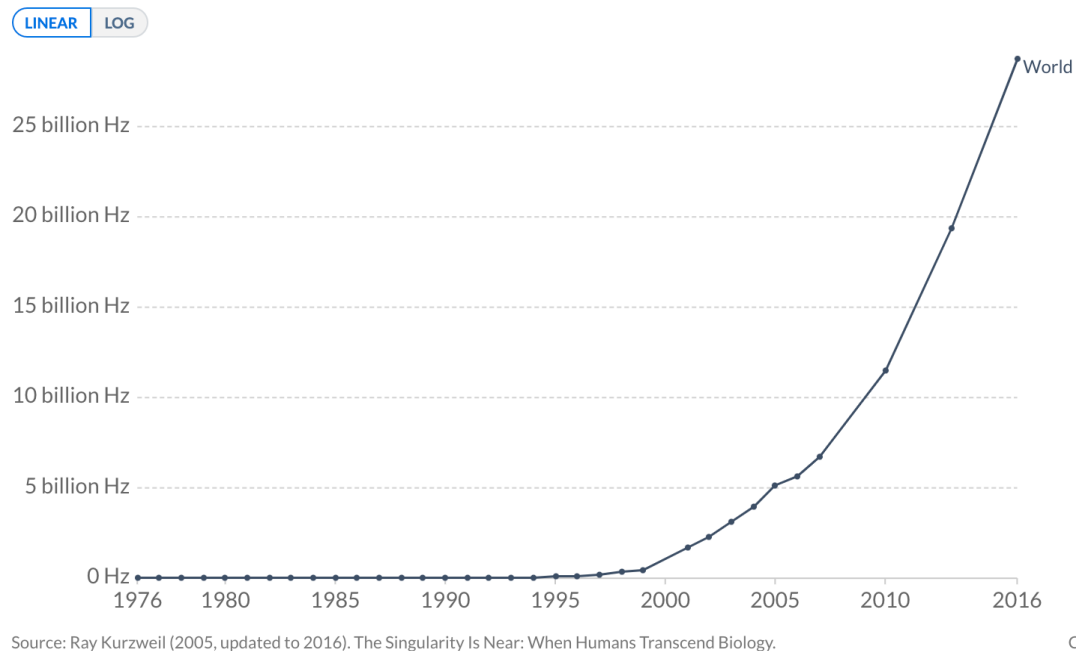


Figure 25 – This linear plot shows microprocessor power measured in clock speed, as the number of pulses

per second (in Hertz) generated by an oscillator that sets the tempo for the processor. (Source: Top500 Supercomputer Database).

3 _ *Cognitive architectures* - Presently, the majority of ANI research approaches of cognitive models are based on how the human brain works though there is still not a generally accepted theory of intelligence, hence these architectures might not follow through.

4 _ *Mapping the brain* - As discussed in the previous chapter on WBE, mapping the brain through reverse engineering will unravel mechanisms to ultimately uncover the whole functional connectome. That process, combined with other such feats, works as a catalyst en route to an intelligence explosion.

5 _ *More hardware* - Moore's Law has been widely covered in this thesis, but even if it would not apply much longer, i.e. even if it would slow down, hardware is still going to grow (*even if not exponentially*). And it does contribute to machine intelligence: "Powerful hardware may improve performance simply by allowing existing "brute force" solutions to run faster." (Moravec, 1976, p). Carbon nanotubes are also expected to greatly improve hardware. It also has an indirect effect by enabling production and processing of large datasets also thereby expanding the IT industry and even the quality of its researchers. (Shulman & Sandberg, 2009). Also, and although the human brain has an astronomical 85 to 100 billion neurons on average (and even a larger number of synapses), there is of course an evolutionary limitation on the volume and the metabolism of the brain that does not find parallel in machines. Hence there is no physical limitation for hardware to grow in the sense that it can always increment using scalable resources. (Muelhauser and Salamon, 2012).

6 _ *Better algorithms* - As an example of improved algorithms without even increasing hardware, Deep Blue computer used 1.5 trillion instructions per second (TIPS) to beat Kasparov in 1997. Only six years later (2003) the program Deep Junior did it with only 0.015 TIPS. Thus, computational efficiency of these algorithms increased a hundred fold in a very short period of time. (Richards and Shaw 2004).

7 _ *Increased serial depth* - The human brain relies on massive parallelization due to the low firing speed of neurons and is therefore not able to perform over 100 sequential operations (Feldman and Ballard, 1982), whereas some algorithms need more

computational resources for parallelization because of its overhead. Hence serial algorithms might provide an advantage. (Muelhauser and Salamon, 2012).

8 _ *Massive datasets* - Various technologies are expected to override Kryder's law (magnetic disk storage density doubles every 18 months) and it is believable that datasets will increase a lot in the near future. (Kryder and Kim, 2009; Walter, 2005). Data availability has been growing exponentially and is expected to continue as can be seen on the plot below. With the increasing amount of data, AI agents can learn more and understand human capabilities better.

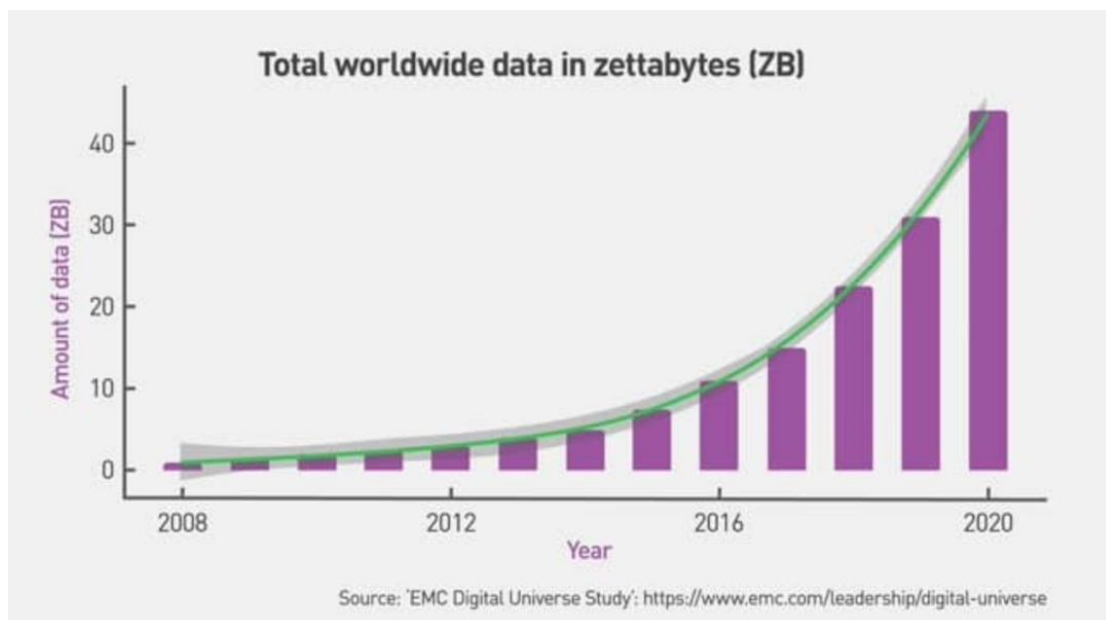


Figure 26 – Exponential growth of dataset in the few years. (Source: <https://research.aimultiple.com/ai/>).

9 _ *Accelerated science* - New tools can accelerate specific fields, as shown in the *Analytical Inferences* section of the WBE chapter, generating a positive feedback. Brain-computer interfaces allowing direct neural access to large databases (Groß, 2009) as Musk's Neuralink example given above, might become regular.

10 _ *Economic & political incentive* - As ANI approaches (or even surpasses) our capabilities in many fields, political (and private) actors that recognize AI's progress may contribute substantial resources in order to accelerate AI's development leading to a kind of AI arms race. (Koza, 2010; Gubrud, 1997). Forrester (a market research company) reported that investment in AI might increase over 300% from 2016 to 2017.

Actually investment in AI startups literally triggered off a few years ago as can be seen on the figure presented below.

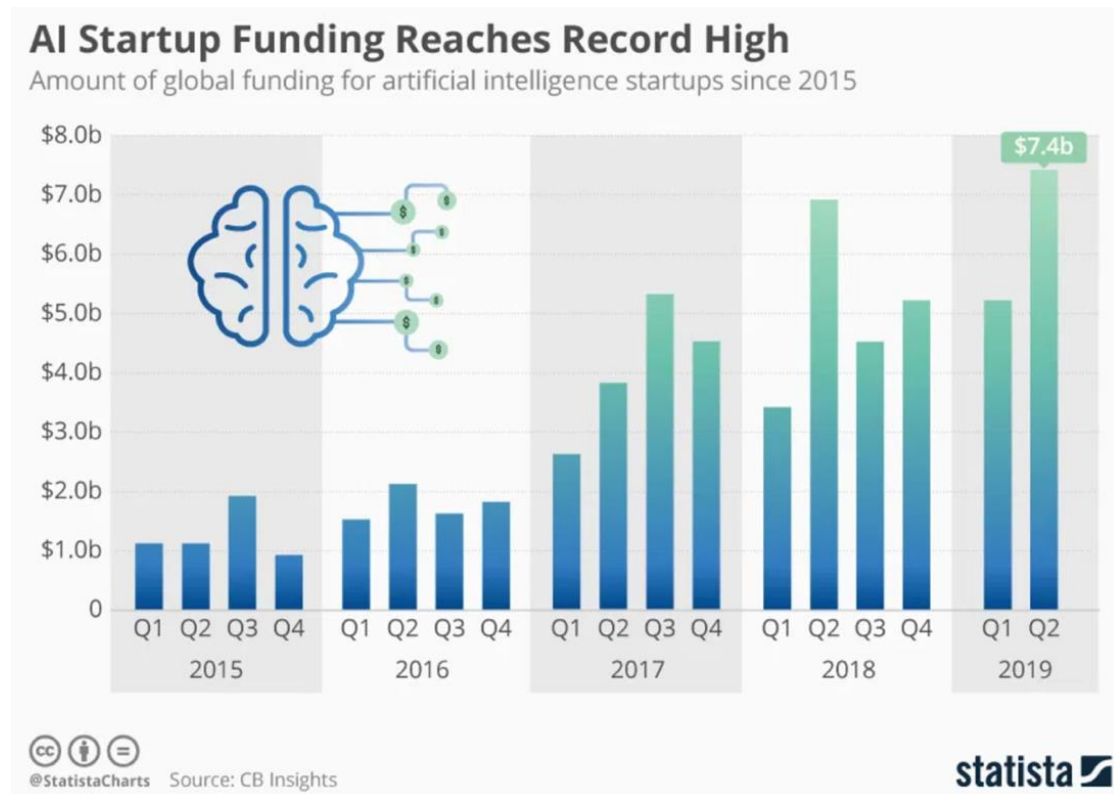


Figure 27 – Venture capital/ startup funding growth in the last few years. (Source: statista charts, 2020).

11 _ Progress in neuroscience - Neural networks and reinforcement learning are methods that have been successfully imported from neuroscience and led to significant AI progress. This trend is expected to continue or even accelerate. (Van der Velde, 2010; Schierwagen, 2011; de Garis et al, 2010). Furthermore, cognitive scientists generally speaking have been uncovering many of the brain's algorithms that contribute to human intelligence. (Trappenberg 2009; Ashby and Helie 2011). Reinforcement learning (RL) is commonly used today in robotics, chatbots and even in programs such as the AlphaGo (that beat humans in Go quite recently). It is a method of learning through experience and improve performance during the learning process itself, which is a common way for humans to learn.

12 _ Duplicability and Editability - An AI's ability can be immediately duplicated just by copying software, which also means that a *digital-mind* population might rapidly expand only depending on its hardware constraints, whereas a biological-mind population

cannot. Thus a turnover of newly skilled AI's might quickly dominate the digital-mind population as better AI's replace others across the hardware base. (Muelhauser and Salamon, 2012).

13 _ *Goal coordination and convergence* - The set of AI's, if coordinated with shared convergent instrumental goals, might effectively objectify any final goals and therefore be capable of overwhelming the optimization power of humankind as a whole. (Omohundro, 2008; Muelhauser and Salamon, 2012; Bostrom, 2014).

5.3 _ The AI Predictions Issue (and (un)reliable experts)

Some of these most promising technologies for producing self-improving software are genetic algorithms and neural networks. AIs will want to preserve themselves to be able to execute their final goals, so instrumental goals are intermediaries *they will want* to pursue. An AI will seek to improve its own rationality because this will then improve its decision-making that will ultimately also improve its ability to reach its goals. It will also seek to acquire the most possible resources in order to satisfy its instrumental goals. (Omohundro, 2008; Bostrom, 2014). This will be further addressed in a section below regarding AI's safety since it obviously impacts deeply on humankind's existential risks debate. Many attempts in predicting ANI's upgrading into AGI, or of an intelligence explosion for that matter, have been made, but there are still no reliably demonstrated methods in forecasting although new methods are being explored (Nagy et al, 2010). Nevertheless the above listed features' progress should be considered individually when predicting ANI's progress instead of the progress of AI/year as a whole. However, given some predictive overconfidence regarding AI's progress in the past (Crevier, 1993), and accounting for potential speed bumps, wide error bars should be wisely applied in this forecasting. There is some evidence of AI recently having made some *previously unexpected* breakthroughs but only few experts expect AGI to emerge within the next decade or two, whereas the majority place their bets around 2050 (*but not necessarily because of Kurzweil's 2045 timeline projection*). (Muelhauser and Salamon, 2012).

In 1950, Alan Turing estimated a 30% pass rate on the Turing Test by 2000. All predictions are constraints on expectations about the future, whether timelines that give dates of specific milestones or scenarios of given conditions that if met will follow to

certain outcomes. A conclusion might be reached about the ultimate state of a specific scenario, given certain conditions about it, in causal models of prediction. Different inputs render different outcomes. Yet when the underlying causes are not well understood, the so-called non-causal models can also help predictions by extrapolating trends into the future as is the case of Moore's Law. Expert judgement, however, might be the best tool in highly speculative fields such as AI, since empirical proofs are not usually available and experts are familiar with common pitfalls. Armstrong, Sotala and ÓhÉigearthaigh argue that since timeline predictions are frequently exclusively based on the expert's judgement, one should decompose the prediction to underline what is evidence-based, what is logical argument and where are the exact points one has to rely on the expert's judgement. (Armstrong, Sotala and ÓhÉigearthaigh, 2012).

Wang and Potter assembled a database of 257 AI predictions from the period of 1950 to 2012. As can be seen on the plot below, the timeline predictions are *randomly* scattered over a large timespan and do not necessarily point to a convergence. Interestingly, there is little difference between the experts' predictions with a standard deviation of 26 years and the non-experts' predictions (standard deviation of 27 years). However, whether we are talking about an expert's prediction or not, there seems to be a trend in predicting AI in a window of 15 to 25 years starting from when the prediction was made. Armstrong and ÓhÉigearthaigh argue that this shows timeline AI predictions to be relatively unreliable. Model-based predictions (causal models and non-causal models) seem to be consistently more grounded than those intuition-based predictions made by experts. (Armstrong and ÓhÉigearthaigh, 2012).

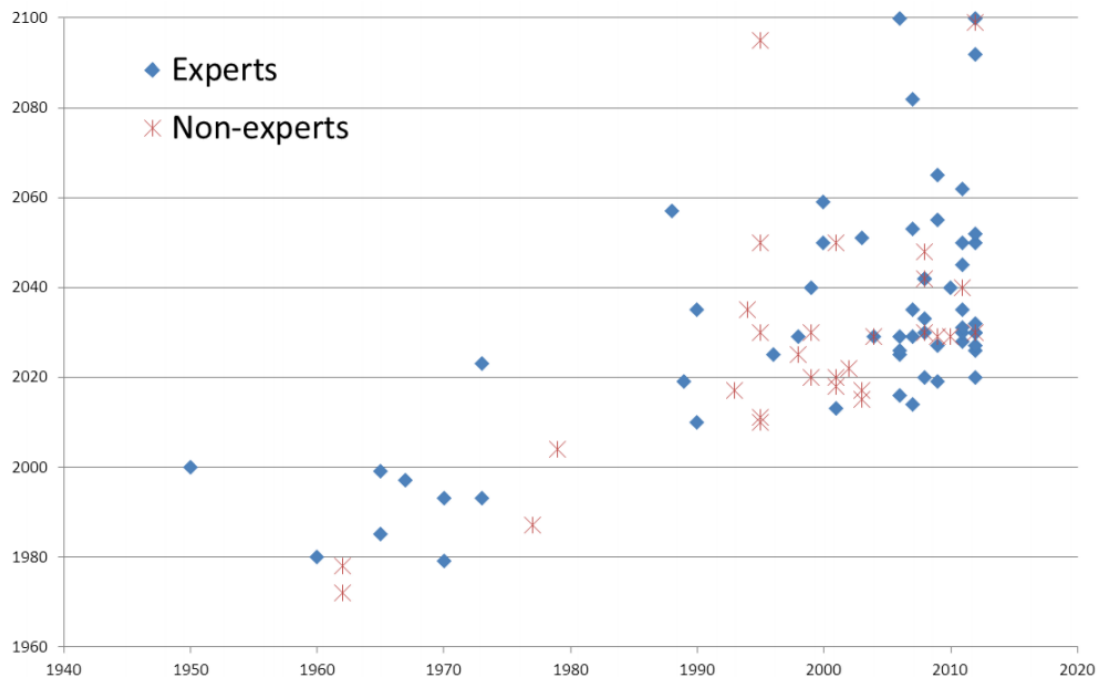


Figure 28 – Median estimate for human-level AI, graphed against date of prediction. (Source: Wang and Potter, MIRI, 2012.)

In the beginning Dartmouth created AI (1956), but also the hype around it and made some bad predictions and this has led to some skepticism regarding AI's progress ever since. If we then add the fact that these were made by some of the major AI experts at the time (John McCarthy, Marvin Minsky and Claude Shannon) and that a couple of decades later the AI Winter took place - it seems reasonable to accept some of that skepticism as a *sensible feeling*. Predictions must be assessed as far as possible, in the context of the time they were made, avoiding hindsight bias (Fischhoff, 1975). Still early at that time (1965), a famous critic of AI (Hubert Dreyfus) began publishing a number of papers downplaying AI's optimistic claims and although not being an expert in AI nor in consciousness made a valid prediction concerning the AI hype cycle: initial progress, grand claims, then unforeseen difficulties and ultimately disappointment. Some of his criticism is also here in this chapter addressed further ahead and it mainly revolves around the importance of unconscious processes, human language and syntax and the problems computers would have in dealing with these. However, he also made some overconfident statements in that critique implying computers would not be able to beat humans in chess (Deep Blue did it in 1997) nor would they be able to cope with linguistically ambiguous questions (Watson won in Jeopardy in 2011). Kurzweil is one of

the most quoted AI predictors and rates himself higher than outside analysts do: Kurzweil commissioned investigations to his own performance which rated him an accuracy of 86% (127 out of 147 predictions were true) and an astonishing 94% (102 out of 108 predictions were reasonably true). Armstrong and ÓhÉigeartaigh argue that having wrong predictions considered as not entirely wrong withdraws some of Kurzweil's reliability, though they do concede that his impressive prediction record even for outside analysts has to be highlighted especially because of the high specificity of some of those predictions he made. It is easy for predictors to be overconfident in their assessments and it is hard to evaluate the predictions' quality at the time they were made. Nevertheless, the potential risks of an ASI mean that correct predictions are important. (Armstrong and ÓhÉigeartaigh, 2012).

6 _ Intelligence Explosion

Irving Good's definition dates back to 1965 but still best resumes the concept of intelligence explosion: A point in time when a machine recursively self-improves leading to a loop of ultra-intelligent machines designing even better machines and then leading to a so-called superintelligence. Once humankind has built an AI with an ability to design better AI's without human intervention, the instrumental goal for self-improvement mentioned above might then ignite a positive feedback of self-enhancement. David Chalmers has also elegantly rephrased this by claiming that when the machine intelligence improves itself, it improves the intelligence that does the improving. Thus, as many others have been claiming since Good's *classic statement*, that ultimate machine, that AI, would literally be humankind's last invention. The following cascade of events would almost certainly lead to a superintelligence. (Chalmers, 2010; Pohl, 2015).

Muelhauser and Salamon argue about an alternative route that usually doesn't catch our eye because of terms such as *self-improvement* and *recursive intelligence that improves itself*: "The machine intelligence could conceivably edit its own code while it is running (Schmidhuber 2007; Schaul and Schmidhuber 2010), but it could also create new intelligences that run independently. Alternatively, several AIs (perhaps including WBEs) could work together to design the next generation of AIs. Intelligence explosion could come about through "self"-improvement or through other-AI improvement." (Muelhauser and Salamon, 2012, p14). As soon as this self-improvement (or/and other AI-

improvement) would take place another question would arise depending on the speed of this take-off, which then in turn also would raise the question regarding the nature of that intelligence explosion - whether it would result in a stable equilibrium made of an ecology of superintelligences or else, if it could derail into a singleton kind of scenario. (Yudkowski, 2008; Loosemore and Goertzel, 2011; Bostrom, 2014). Either way, such an intelligence would surpass us humans as we did with the apes. This analogy is frequently made but it does not really pay tribute to that difference since it took some time for humankind to detach from those other primates which means that no single ape actually witnessed that phenomenon. However in this case it would be different, certainly if this would be a hard take-off happening very quickly as in days, hours or even seconds, but also even in the other intelligence explosion scenario that would not take more than maybe years or even decades, but surely not more than one single human person's lifetime. As such, one single human being (or even every living human being) would witness it. Also, these *super AIs* might become motivated in preserving their initial goals which on the other hand would also mean that ever more capable AIs would eventually be also ever more efficient in pursuing their goals. Hence the main question regarding an intelligence explosion of any kind would be that of making sure those AIs would have to entail human values for they might be endangered in a purely incidental way or even by pure negligence. (Yudkowski, 2008).

The type of intelligence explosion, as well as the intelligence of both – our intelligence and particularly the one responsible for the singularity - will most surely determine the kind of perception of the singularity itself. How would we as human beings perceive such an event? In the so-called *inward explosion*, as Hutter claims, “where a fixed amount of matter is transformed into increasingly efficient computers until it becomes computronium” (Hutter, 2012, p 6), it is debatable whether we would be able to connect the speed of affairs taking place in a singularity, and therefore we would probably not be able to understand it. Communication between us and the virtual population would over time become problematic, and this considering it would still take some time, i.e, not necessarily in case of a hard take-off. Considering the self-recursive improving kind of nature of that virtual population, and unless our technology would by then be as advanced as to cope with this “explosion”, it seems unlikely we would comprehend the state of affairs as they would unfold. When observed from the outside (us humans), a society (or even a system) of increasing intelligence would eventually over time become increasingly indistinguishable from noise. (Wagner, 2017). On the other hand, in the case

of a so-called *outward explosion*, an increased amount of matter being transformed into computers of fixed efficiency would lead to a resource competition between the virtual population (the self-recursive system, aka AI) and the observers (humankind) of the singularity, which would also account for an increment of the existential risk. (Hutter, 2012; Wagner, 2012).

7 _ AI Safety (and risks of The Singularity)

There is a kind of popular hype raised by dubious cinematic artworks that commonly links the singularity to an army of hostile superintelligent robots as mentioned with *the terminator* example at the beginning of this chapter. This singularity meme, a cultural construct, has unprecedentedly tied the singularity to malevolence, which actually is more of a human kind of feeling, rather than an AI's characteristic (Dawkins, 2000). Critics of the singularity often mix the singularity hypothesis with this pulp concept of the singularity meme, reinforcing the confusion already in motion. However, the majority of *the singularity* proponents distance themselves from the meme and explicitly reject the apocalyptic robot uprising scenarios. Authors of scientifically or philosophically *sound* singularity hypotheses underline that instead of malevolence one should be talking of indifference, inasmuch as this difference really distinguishes the scientific hypotheses from the meme. In fact, and just as a side note - these apocalyptic fantasies involving alleged singularities resulting in malevolent AIs are to large extent even mistakenly called science-fiction. The hostile AIs' uprising scenarios can be considered if not implausible views, then at least very much anthropomorphic and humanly biased. (Eden et al, 2012). It seems more appropriate to approach these scenarios with the value of indifference rather than malevolence in mind, also if we consider self-preservation as a basic drive of AI (Omohundro, 2008); or an emergent property of superintelligence rather than being explicitly programmed (Bostrom, 2014): it seems highly improbable that an AI would spontaneously exhibit specific human traits. Unless these were programmed or ingrained in it as part of some kind of safety strategy. In conclusion, we have to think of singularities bearing the traits of indifference rather than malevolence when approaching the associated risks.

Furthermore, rather than AIs suddenly undergoing unexplained emotions or needs as the recurring anthropomorphic bias would sustain, an existential risk may actually be the

fact that they would lack these values instead of emphatically exhibit them as so dramatically happens in those *sci-fi-ish* kinds of scenarios. Actually if we take ourselves as an example towards other less intelligent species, we notice that it is indifference that defines us too, rather than malevolence of some kind. It would also be more expected from a superintelligence to act indifferent toward us, which does not mean it would be harmless, just as we are not toward the less intelligent species. This of course means that, and as was duly pointed out by some of the interviewees (see SSI-Analysis, Part V, Question I-Q2 - Barrat and Sandberg), they could really represent an existential risk to humankind just by pure indifference. Bostrom's Paperclip thought experiment is a good example of the danger coming out of a superintelligence just by indifference. This indifference resides in the lack of humankind's core values ingrained in the AI, which would make that AI an existential risk to humankind. (Yudkowski, 2013; Bostrom, 2014).

If an AI would lack friendliness or humanlike philanthropic kind of values built in its goals, then the aforementioned scenario could become true and end humankind just as a byproduct, as claimed by Bostrom: "The risks in developing superintelligence include the risk of failure to give it the super goal of philanthropy. One way in which this could happen is (...) that a well-meaning team of programmers make a big mistake in designing its goal system. This could result (...) in a superintelligence whose top goal is the manufacturing of paper clips, with the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities." (Bostrom, 2003, p 12-17). In another example of misguidance and lack of control, these super-AIs could without the right set of values programmed lead humankind to a false utopia just by carrying out the goal of making humankind happy, for instance. Of course the misguidance or even the misuse in AI as in any other technology is not an act of intentionality, though terrorist affairs might find a gold-mine in developing ASI.

The Future of Life Institute recently published an open letter regarding AI safety: "Just as most chemists and biologists have no interest in building chemical or biological weapons, most AI researchers have no interest in building AI weapons." (Future of Life Institute, 2015). Yet the military potential of AI (let alone of ASI) is as big (if not bigger) as the compromise between nuclear clean energy and the production of atomic weapons, which led the Future of Life's Institute to warn against an eventual AI's arms race: "Unlike nuclear weapons, [autonomous killer robots] require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military

powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc.” (Future of Life Institute, 2015). At the 2017 Asilomar conference, *The Future of Life Institute*, acknowledging the momentum, and to downplay negativism over AGI, set out principles for its ethical development in a kind of manifesto or guideline. (Gans, 2018).

It seems clear that the emergence of self-improving AI may be dangerous which is why building a stable AI with a *philanthropic utility function*, aka a friendly AI, is so important. (Yudkowski, 2001). Now, what exactly is a philanthropic utility function? And to what extent is it not directly proportional to the AI’s intelligence?

7.1 _ Strategies to control (artificial) intelligence

As already mentioned earlier, *Universal Intelligence* is the metric for measuring the intelligence level of any kind of agent in a specific given environment. (Legg & Hutter, 2007). Yet, its applicability has not been tested empirically and it defines the intelligence of that agent as a function of the environment where it operates. This means intelligence would in this definition be proportional to the agent’s ability in solving problems that specific environment would pose, which of course also would mean that agent’s ability in solving other environments’ problems might be very low, hence not exhibiting intelligence at all (in those other environments). The difficulty in defining intelligence undermines our ability in assessing its risks: we cannot foresee what capabilities an intelligence different from ours would be able to develop and what applications it would find for it; we cannot foresee the likelihood of losing control over technology; and we cannot foresee the consequences of an emergent future intelligent technology: controlling the behaviour of agents more intelligent than their makers is difficult if not impossible (Bostrom 2014) — in particular if the agents learn to modify themselves. (Yampolskiy 2012).

The main strategy *to control intelligence* has been to develop (or try to) friendly AI for example by hardwiring artificial agents with our core values, which again raises another question: What exactly are our moral values? (Yudkowski, 2001, 2013). The concept of a friendly AI very much depends on an objective theory of morality, or at least of some kind of universal concept of “good”, which is clearly a difficult goal for computers (or AI)

to achieve (*even difficult for humans*). (Brundage, 2014). Also, and even assuming a behavior such as an objective friendliness could be established, software engineering seems to show that implementing desired behaviors is far from easy. (Wing, 1990; Hinchey et al, 2008).

Another desired strategy for trying to control AI is a sort of confinement, in order to keep the AI inside a specific environment for example through special hardware. This is known as *the AI boxing approach*, which of course also has its weakness as Vernor Vinge alerted: In the case of a superintelligence this would hardly be effective since it would easily come up with something to free itself, even something like tricking humans on purpose with some kind of advice that would accidentally set it free. (Vinge, 1993). Another such strategy is *the Oracle* presented by Bostrom. The Oracle would be an agent always choosing yes and no answers therefore minimizing AIs capabilities by limiting them to only one possible action, which would downplay its potential impact. (Bostrom, 2014). However, and as alerted by Vinge, if superintelligence should in fact be of a dimension several orders of magnitude higher than ours, would it not be logical to expect it to find itself out of that *boxing confinement*? If not, then one might suspect its intelligence not to be that super after all. An unfriendly SI could deceive humans by devaluing its own intentions and abilities. Thus, and again as discussed with the interviewees (see [SSI-Critical Analysis](#), Part V, [Question I – Analysis #2](#)), holding strong to the probability of technology always ending up on the upper side and as concluded in the analysis of that same question: „With AI as with other technologies, we may become victims of “the tendency of technological advance to outpace the social control of technology. (Posner, 2004).” (Muelhauser and Salamon, 2012, p32). On the other hand, trying to ban its progress would most certainly lead to its underground research which could turn out to be even a worse scenario as it has been suggested (Joy, 2000), hence research for a *friendly AI* has been gaining a lot of traction these last years. Goertzel even proposed a kind of *Nanny AI*, that would postpone chances of AIs’ self-improvement therefore also forestalling a singularity giving time to safely build a kind of benevolent or beneficiary singularity (Goertzel, 2012), since we might not get a second chance as Yudkowsky claims. (Yudkowski, 2008). Also Bostrom has proposed a program of *differential technological development*: “to retard the implementation of dangerous technologies and accelerate implementation of beneficial technologies, especially those that ameliorate the hazards posed by other technologies.” (Bostrom, 2002, p18). In conclusion, in order to avoid unsafe AI development, there have been approaches suggested to control an

intelligence explosion through internal and external constraints for goal-directed AIs via physical and software confinements, but none of them have until now been truly satisfactory. (Chalmers, 2010; Yampolski, 2012). There are essentially three reasons why superintelligence might represent an existential risk in Bostrom's perspective:

1 - The likelihood of an emergence of a singleton. The first developed AI reaching such a status (superintelligence) would probably have a competitive edge over all others (including humankind, of course);

2 - The orthogonality thesis: "Intelligence and final goals are orthogonal (...) more or less any level of intelligence could in principle be combined with more or less any final goal." (Bostrom, 2014, p107). He claims such a singleton might not have human values and to sustain this hypothesis recurs to the ability in actualising abstract goals as the metric for intelligence. A kind of means-ends analysis;

3 - Convergent instrumental goals (*which Omohundro calls basic AI drives*): In the paperclip factory example, if the AI had the final goal of converting everything into paperclips, it would most probably actualise its specific convergent instrumental goals eliminating potential threats and acquire all possible resources in order to facilitate that final goal (in this case making paper clips).

Omohundro argues that sophisticated AI systems are likely to exhibit tendencies that will be present unless specifically counteracted, and calls them basic drives. (Omohundro, 2008). Barrat talked about these drives in his interview for this thesis (SSI-Analysis, Part V, Question II-Q2) in reference to the thought experiment *iterating human intelligence* discussed at that moment, and although complimenting the evolutionary algorithms approach altogether, these basic drives would however sustain on the presumption of *how they would act*, thus, in Barrat's view, a doubtful approach.

A kind of biomimetic approach to tackle the existential risk issue, that I would (perhaps a bit pretentiously) like to call *bit-apoptosis*, could also be a strategy to consider. In living systems, and given certain adverse conditions, cells commit a kind of suicide to keep the systems' regular functioning. Assuming an AI would delegate specific AIs to targeted goals, this *apoptosis concept* could also be applied to artificial intelligence systems as soon as an ANI software would reach a specific pre-AGI stage, an apoptotic component could send it automatically back to an ANI stage, thus avoiding an intelligence explosion and all the rest. In the paperclip scenario, the AGI could self-regulate against a potential

apocalypse by not switching on specific AIs. (Gans, 2018). However, and independently of how exactly this apoptosis would be reached, whether through some kind of threshold in power accumulation, or through a well-devised algorithm that would force the AGI to take that choice, ultimately, and when or if specifically referring to superintelligence - it would again mean a way to try to outsmart the ASI. By default, and once more, this does not seem very probable, just as an ant cannot deliberately outsmart a human being. Thus instead of trying to outsmart ASI, the strongest bet remains in trying to solve *the control problem*.

Given these plausible outcomes and to solve *the control problem* and develop a safe AI, Bostrom has suggested another strategy for minimising existential risk: *Indirect Normativity*. Programming the seed AI to learn about *friendliness*, instead of trying to input it directly into its motivations or goals would make AI the engineer of its own values, which would then hopefully also off-load itself its final goals. One can call this approach a kind of switching from direct coding for friendliness to indirect normativity. A reliable friendly kind of super AI could then develop even if we, as humans, do not know exactly *what that friendliness is*. (Bostrom, 2014). Again, this urges back the question about our moral theories and as stressed earlier in this chapter, it emphasizes the contribution moral philosophers can have in designing values to ensure AI safety given that the only reference superintelligence might have regarding ethics without this specific input would be the history of humankind's moral philosophy.

As a final note, it is clear that the AI safety problem can be divided into two main approaches to these risks: One that tries to limit the system's capabilities, in particular of control, and another that tries to input our own human values in the system. In the best case scenario these intelligent systems would learn those values by themselves (we would only sort of ignite them), and as Omohundro suggested and was also mentioned by Barrat in his interview for this thesis (SSI-Analysis, Part V, Question II-Q2 the systems would kind of compete, always keeping each other in check. Thus no particular agent would gain a strategic lead over others and allocate resources to safely generate an ASI, hence minimising future risks.

Now, this may seem implausible or at least a mountain of complex issues to be successfully tackled. However, what seems implausible to us humans might not necessarily be *objectively* implausible and as Joel Thomas claims in his critical review of Bostrom's book, *Superintelligence*: "Although this reasoning could be used to justify any

argument, it becomes relevant when creating inconceivably powerful beings that have a high chance of having fundamentally different conceptual systems that we do.” (Thomas, 2016, p1093). Furthermore, and as mentioned earlier, it is not mandatory that *general intelligence* has to be generated through the same evolutionary processes that produced human intelligence. (Muelhauser, 2012; Hutter, 2012, Bostrom, 2014). *Actually, evolution itself managed to create intelligence without intelligence to do so.*

8 _ Politics & Skepticism

Superintelligence is, as was the case described for MU in the previous chapter, a technological achievement of such an incommensurable magnitude that its outcome might by default change everything humankind knows of. On the other hand, humankind’s existential risk associated with superintelligence’s disruptive power, is for many experts in the direct proportional dimension of its power. This alone is fertile ground for skepticism and reason enough to assemble a legion of skeptics.

Today, most of the research and development in the field of AI is still mainly focused on narrow AI (ANI). Many experts (and not only) fear another AI winter that would hold back AI’s general progress, which ironically (*or not*) makes computer scientists the first cluster of SI skeptics. Actually as soon as Irving Good in 1965 first started talking of an intelligence explosion eventually leading to more than human intelligent machines, skeptics began to express their critique. Dreyfus, also in 1965, critiqued AI’s chances of eventually matching human’s intelligence recurring to Descartes’ famous *Discourse On Method*: “it must be morally impossible that there should exist in any machine a diversity of organs sufficient to enable it to act in all the occurrences of life, in the way in which our reason enables us to act. (Descartes, 1637)” (Howard, 2000, p45). The *improbability* of actually building a superintelligence based on computational theory is one of the main source of skepticism. The sceptics argue that SI cannot be built because it would require an advanced kind of computing that cannot be made by us today, nor by the existing AI. Others argue that it might be possible to build such advanced systems but not within the next decade or two, hence worrying about it now would be premature and it therefore is not worthy of current attention. (Pohl, 2017; Baum, 2018).

There is also a kind of political skepticism (*risk-profit politicized skepticism*) that aims to protect corporate interests or profits and promote anti-regulatory ideals, thus leaving

major corporations with enough room to develop superintelligence without being regulated. Baum argues that these financial and ideological motivations are not about science, they are rather political and that science is ultimately manipulated towards those goals. Also, the fact that there already exists a lot of skepticism about superintelligence only ends up helping out the cause. Moreover, scientists usually tend to be very cautious and generally avoid statements before being absolutely sure of them which not only helps them raise or maintain vital funding as also keeping off from speculative discussions ultimately grants them valuable credibility. On the other hand, this also ends up reinforcing the appearance of uncertainty of that specific science or topic, in this case, discredits the eventual emergence of an AGI or even ASI, which again helps out that *politicized cause*. Hence skepticism can be used as an effective tool to promote certain political goals. (Baum, 2018).

Baum argues that although it is still relatively soon to talk about political skepticism in regard to superintelligence there are some signs such as AI academics downplaying that hype in order to protect their field's reputation and avoid criticism prompting in cutting fundings, which already happened in the past with the *AI Winter*. The possibility of another AI Winter actually works in favour of this politicized skepticism. Facebook's influencing elections case, although not related to superintelligence (*at least not yet and/or not publicly*) is an example of this in the sense that its associated risks had to be downplayed in order to avoid further regulation. Major corporations such as Google and Amazon are of course also some of the major stakeholders in ANI, AGI and potentially ASI. Sandberg and Barrat discussed in the interviews for this thesis (SSI-Analysis, Part V, Question II-Q2) a number of scenarios regarding these major corporations' power in the development of such an intelligence inasmuch as eventually even becoming the first collective superintelligences in a not so distant future. Barrat even discussed their doubtful ethical motivations as corporations clearly hinting at a political stance they have and at a political stance they might achieve with the development of such AIs. Political skepticism' examples in our societies' history can be found in the tobacco industry, in the fuel and oil industry and even in the pharma-health industry. Yet, contrary to these corporations, IT companies such as the ones discussed in these interviews, not only have the assets to conduct these policies, but also the power to operate in a privileged niche, as the Facebook example showed.

9 _ Concluding Remarks

Most of the authors agree that in order to reach an intelligence explosion and develop a superintelligence, there are essentially two major requisites that have to be reached: firstly, a powerful hardware system in storage capacity, in processing capacity and in speed. All these seem to be on track. Speed and capacity keep doubling every year and cloud computing permits large clusters of processors to be created dynamically. Secondly, AGI software will have to be self-improving and self-aware (*if not for other reasons, then for its survival*). Major opposition to these optimistic claims finds support in arguments claiming it is impossible for AGI to be achieved through research approaches based on logic combined with neural networks. (Dreyfus, 1997). However, as Pohl argues in defence of AI's recent progress: "Nobody could have predicted a global knowledgebase (*Internet*), enormous networked computing power (*Cloud computing*), and the promise of a hybrid bottom-up (*neural networks*) and top-down (*symbolic representation and reasoning*) approach." (Pohl, 2017. p10). Today, it is becoming generally accepted among AI experts that artificial general intelligence and its subsequent superintelligence might be a reality someday, and maybe not that far away as an informal survey made at the 2011 AGI Conference in California suggests: 42% claim it will be a reality by 2030; by 2050 (25%); by 2100 (20%); by 2150 (10%); and, never (2%). (Barrat, 2013, 196-7). As such, there seems to be almost a consensus that it will come about approximately around mid-century.

The core ideas of the *mainstream* of these experts from the Machine Intelligence Research Institute such as Yudkowski and of the Future of Humanity Institute such as Bostrom are summed up as follows: according to known physics, SI is possible through self-modifying software. SI's emergence will be hard to predict but an AGI with human-friendly goals able to self-modify might develop itself into a condition where it no longer pursues the same goals it started out with. Hence, self-modifying SI might probably diverge from human values and it would then have little to none motives to care about humans (*just as humans don't care about ants' affairs*). Thus, the existential risk to humankind.

9.1 _ Conflicting views and different approaches I

Bostrom's book *Superintelligence* essentially asserts this existential risk posed by superintelligence and advocates that its development has to be regulated. Goertzel criticizes that Bostrom's major ideas have actually already been suggested or proposed by others such as Yudkowski and Omohundro, yet not in such an elegantly philosophical style as Bostrom's. (Goertzel, 2016).

According to Omohundro, self-improving systems *will* be made of primary drives such as: Efficiency - to optimize available resources; Self-Preservation - to avoid being turned off (driven by their goals); and Resource Acquisition - to achieve their goals regardless of the impact they will be compelled to gather all resources they need. (Omohundro, 2008). As mentioned earlier in the AI's safety section of this chapter, Bostrom's *instrumental convergence thesis*, and values for that matter, also argues that these intelligent systems will focus on their "goal content" and on the acquisition of relevant resources. Both Bostrom and Yudkowski argue that these systems with a self-preservation drive will be more likely to survive (*some resemblance to Dawkins' selfish gene, but applied to AI?*). Goertzel counters that self-preservation is a biomorphic idea, or even anthropomorphic and argues that these systems, being superintelligent, might go far beyond such notions. He argues that in competitive scenarios, systems that would develop new different systems from themselves might actually be the elected ones. Furthermore, Bostrom and Yudkowsky argue that these systems would be indifferent to human values and thus more likely to do bad things to humankind. Yet this likelihood is arguable as Goertzel claims: "Uncertainty doesn't merit assignment of a high probability to the outcome!" (Goertzel, 2016, p67). Actually, all that uncertainty does is lower any kind of prediction, thus not necessarily with a probability of favouring more scenario A instead of scenario B (whichever those are).

Goertzel criticizes Bostrom's discourse of presuming that a hard take-off scenario would lead superintelligent systems to do bad things to humankind instead of emphasizing all the positive things it could do. These are pretty much diluted (*in Bostrom's book Superintelligence*) and do not get as much attention as the dangerous side of SI. Bostrom claims that: "a fast takeoff occurs over some short temporal interval, such as minutes, hours or days. Fast takeoff scenarios offer scant opportunities for humans to deliberate. Nobody even noticed anything unusual before the game was already lost." (Bostrom,

2014, p 252). In fact, assuming it would have to end up in something like *us against them* might be a bit biased and overlooking or dismissing the possibility that it might actually be beneficial for us, this kind of *reasoning* could be tied to our humanlike concept of intelligence.

Omohundro also proposed the “safe-AI-scaffolding” approach, which would require next gen systems to be built on the previous ones with a safety proof in each generation. (Omohundro, 2014). But human beings won't be able to control these systems once AI has reached that threshold, and if in fact superintelligent, it *will* find a way to avoid any kind of confinement or constraints. Thus, communication of any sort with such a non-biological SI might be improbable and furthermore humans might not even be capable of determining its intentions, whichever these might be. Even assuming a slow kind of take-off scenario, would it be possible for humankind to cope with these systems in order to ensure a kind of safe development? It was not possible to obtain a real consensus regarding this issue in the interviews made for this thesis, though as discussed in that section (SSI-Analysis, Part V, [Quesion II](#) + [Analysis #3](#)) some of the experts expressed disbelief in a hard take-off scenario. However, I reiterate that it is likely that once an AI is capable of self-improvement, its progress would perhaps by default be exponential toward an AGI / ASI. And therefore, whether fast or slow take-off, it would never give humankind enough time to react and thus any kind of measures would always have to be prophylactic to be taken before ANI reaches AGI like improving our own intelligence as discussed in the TE-3b (besides friendly AI).

The *virtual confinement*, the *scaffolding approach*, the introduction of *friendly biases* and the *apoptotic approach* all discussed above, are all reasonable approaches for AI-safety but none of them is flawless. Thus, a mix, a kind of cocktail of many of these strategies may ultimately be a way to mitigate the singularity's related risks.

9.2 _ The fear of a Singleton (*and of a non-human consciousness?*)

While it has been argued that an AGI is not required to exhibit self-awareness in order to be capable of self-improvement, it does however require self-awareness for an intelligence to be able to determine its own goals. Self-awareness is the ability for introspection and for one to recognize itself as an individual, as a separate entity, from

other individuals and from the environment. Self-awareness is ultimately a trait that a singleton must possess. It is the recognition of one's conscious affairs. It seems straightforward that self-awareness is to some extent a reflection of one's intelligence, i.e., one needs to have a certain intelligence in order to perform that introspective act of recognition. But how much tied they are together is again matter for speculation and how much intelligence is required to generate such a feature, too. However, and as discussed in previous chapters (WBE & MU), given specific conditions as advocated in functionalism, a consciousness might be generated in a non-biological intelligent system. Now, only that could enable a singleton to be a *singleton*, hence establish its own goals.

Bostrom postulated *the orthogonal theory* to argue that any level of intelligence could be combined with any final goal. He also postulated *the instrumental convergence theory* where sub-goals might be predictable based on the nature of the final goal. Now, I would counter that: 1) first - if the superintelligence is in fact, or has the characteristic quality to be, a singleton, then it is able to rewrite its own source-code, hence alter its final goals whenever it desires (or even *feels like*); but also, 2) since we never dealt with intelligences superior (let alone astronomically superior) to us, we haven't got the faintest idea how it would behave, think, act and be! By definition we cannot measure the outcomes of a singularity, thus how might we expect such an intelligence to act? Quoting Vernor Vinge on this, just as a reminder: “singularity – a place where extrapolation breaks down and new models must be applied – and the world will pass beyond our understanding.” (Vinge, *True Names and Other Dangers*, p. 47).

Is it possible that what we really fear about a singularity / superintelligence, is not actually its intelligence, but its consciousness? Might it be then, that driven by this anthropomorphic bias, we fear a singleton might have a consciousness such as ours, and therefore not only fear ourselves but in fact fear that what we are, which once exponentiated might be something terribly dangerous? In a sense we already do have *singletons* today on Earth and without SI as Sandberg and Hanson also highlighted in their interviews for this thesis (SSI-Analysis, Part V, [Quesion II](#) + [Analysis #3](#)).

9.3 _ Conflicting views and different approaches II

Bostrom advocates in his book *Superintelligence*, the hypothesis of assembling a very tight group of experts to control AGI development to downplay the singularity's existential risks. In his rationale, avoiding an open access kind of project might assure a safe and positive outcome to humankind. (Bostrom, 2014, p251-4). Who would select these people? On what grounds? And to what extent does it make sense to possibly put the world's fate in the hands of an elite? Is this not a political, cultural and socially debatable issue, maybe even as dangerous as the supposedly maleficent superintelligence? Goertzel and Pitt, on the other hand, suggest quite the contrary: To develop an AGI in the open source domain. Of course one can argue it would therefore be more exposed to probable hackers and terrorists (etc...), but it would also surely be a more democratic process, hence also regulated and participated by the whole world. (Goertzel and Pitt, 2012).

What is in fact happening today as mentioned earlier, since ANI's progress is mainly determined by major corporations (Google, Amazon, Baidu, Facebook, Apple, etc...), is that AGI's development is more likely to come out of these elites, than out of one selected and submissive to and by the UN, or governments of the world generally speaking. The OpenCog project led by Goertzel and Pitt aims to diffuse intelligent systems' development in an open-source made of researchers, students, startups (etc...), hence also increasing the variety of projects which ultimately might not just target a more balanced outcome as also an eventual global brain. A kind of *mindplex*. (Goertzel, 2016).

Furthermore, Goertzel argues that this kind of intelligence, contrary to Bostrom's and Yudkowsky's proposals, is actually truly general since it is not conditioned by a predetermined goal: "The theory of open-ended intelligence rejects the idea that real-world intelligent systems are fundamentally based on goals, rewards, or utility functions. Ultimately, the initial goals assigned to an OpenCog system by its programmers may serve as nothing more (or less) than a way of helping the system to get its own, open-ended, self-organizing intelligence off the ground." (Goertzel, 2016, p67). Goertzel's argument of open-ended intelligence is pretty much based on Weaver's *non-reductive* concept of intelligence beyond *optimized utility functions*. This open-ended perspective of intelligence makes it more dynamic in the sense that it can thrive its own goals, and new ones, throughout its development instead of being restricted and reduced to a goal-

oriented system kind of intelligence. Goertzel argues that an intelligent system conceived and driven to pursue its original goals cannot develop.

9.4 _ Final Notes

Koene argues that there are constraints about a SI we should focus on such as the physical requirements it would need to carry out its iterations of self-improvement and to what extent could its advancement be slowed down because of resources it would depend upon. He suggests these kinds of issues are out there for us to intervene in advancement and that we can affect the course of actions in practical terms rather than focus on *mission impossible* kinds of actions such as pursuing friendly AI scenarios or even trying to manage or control everyone working in AI. (Koene, 2013). In this line of thinking, I would make a quite trivial analogy to our current approach toward AI safety and SI's *risky emergence*: We are a bit like the underdog team that knows not to stand a chance against the big *Goliath kind of team*. So we focus on trying to make SI-team's life harder by undermining it and trying to control and take out its best player (intelligence) instead of focusing on our own strengths. Those teams usually (and statistically) end up losing. Intelligence always was our strength and as argued at 4.1 *TE - iterating human intelligence applicability*, we might as well place our bets on ourselves and focus on our own *possible intelligence explosion*. Furthermore, as Koene advocates, a singularity driven by SIMs (substrate independent minds, or *uploads* for that matter) instead of by AI, has by default less risk associated due to our participation and therefore also allows humankind to embrace a competitive universe.

A series of prominent personalities have been warning of AI's arms race as the citation of Stephen Hawking shows right at the beginning of this chapter. A collection of groundbreaking research in SI and brain augmentation related issues has recently won a major new European research prize (2017)*. Led by neuroscientists, more than 600 authors contributed with papers about brain augmentation research which could make us think if we'd just paused for a moment to do it: Only a couple of decades ago this was truly an exclusive science-fiction matter. Before the year 2000 it would have seemed absolutely ludicrous just to envisage such a scenario. Personally I can still remember this time very vividly. I might remember these very lines in another couple of decades to witness a singularity / superintelligence occurrence in real-time. However and as has

been argued in this chapter, predictions are and have been reasonably uncertain. Especially in the case of an emergence of human or superhuman-level artificial general intelligence (AGI). Epistemically reasonable predictions need to smear out their probability distributions pretty much all over the future time axis.

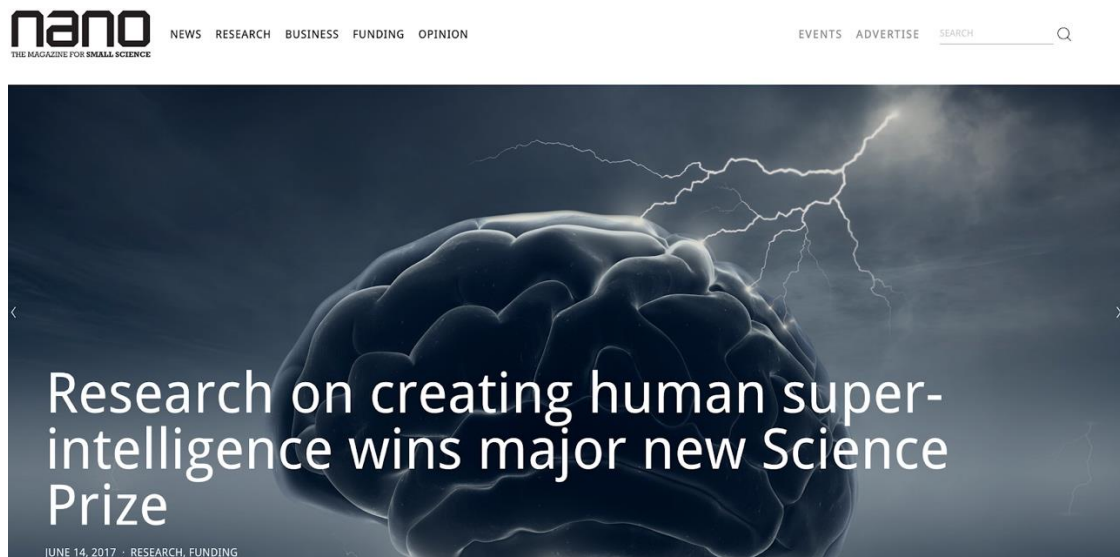
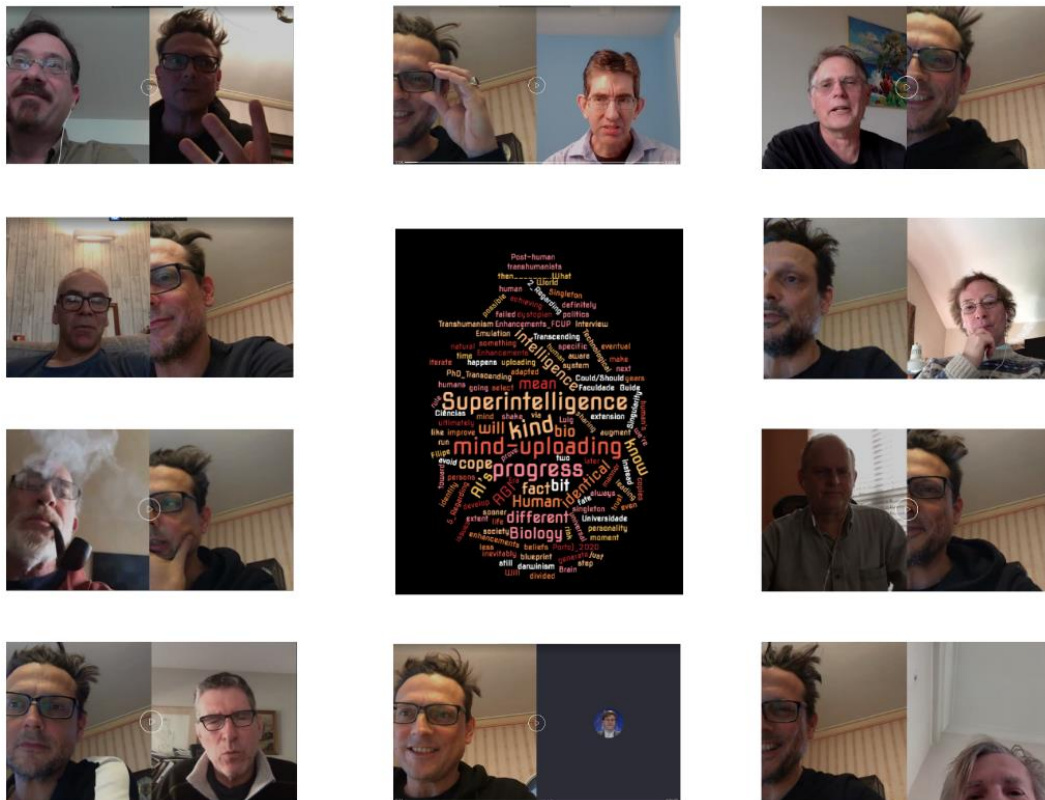


Figure 29 – Screenshot taken from cover of nano-magazine issue June 2017.

PART V _ SSI Analysis



This is the first age that's ever paid much attention to the future, which is a little ironic since we may not have one.

— Arthur C Clarke

1 _ Intro and methodology

This chapter reviews the semi structured interviews to various experts held for the purpose of this thesis to qualitatively contribute to the enrichment of the research undertaken. This is also the chapter where I spell out in more detail the methodology followed in this part of the thesis.

The issues explored in this thesis, such as whole brain emulation, mind uploading and superintelligence, involve a fair share of polemic. Discussions revolving around them require expertise in different areas, such as cognitive science, neuroscience, philosophy and ethics. It must also be mentioned that these possibilities are thinkable – and of course explored by meaningful thought experiments – but that there is no consensus on whether or not they are actually going to happen, or when, or what their (desirable or undesirable) consequences would be. Therefore, these are issues that call not only for a critical appraisal but also for an interdisciplinary exploration that only has to gain if it succeeds in mobilizing a wide scope of different expertises that can help make sense of the technological conditions of possibility and the philosophical, ethical and political implications of these issues.

This is the part of the thesis in which the scientific themes discussed and previously reviewed in this thesis through a careful philosophical and scientific assessment are explored with a different methodology, which draws more heavily from the social sciences. It goes without saying that in this endeavor the aspect of methodological construction (Harreveld et al, 2016; Fortin, 2009) is very important. This being a qualitative research (Carvalho Guerra, 2009; Denzin and Lincoln, 2018), it is devised with a mix of different techniques, the most important of which is the interview (Ketele and Roegiers, 1999) even though the script was predistributed to participants as if it were a questionnaire.

Bearing that in mind, I decided to conduct a series of semi-structured interviews. And it must be stressed how and why this was done. First, the choice of the interview. I decided to conduct semi-structures interviews (Ketele and Rogiers, 1999) because this in the kind of data collection that is the most appropriate to gather deep knowledge about the matters at hand. My primary goal was not to assess the opinions of laypeople on these topics. If it were, I would most likely have devised simple questionnaires with several

closed questions and tried to reach as wide and representative a sample as possible, in order to then explore these answers quantitatively. However, given that my target was to gather *expert* opinions on these cutting-edge scientific issues in order to then include them in my effort of *communication of science*, I decided to constitute a non-probabilistic sample by rational choice, and which meant to be as representative of the most important experts in these topics as possible. For this I contacted a group of experts in different fields, all of them related to these scientific issues, whether through the lens of transhumanism and human enhancement and coming from a more scientific perspective, or through a more philosophical / conceptual perspective. Some of these experts are more optimistic, while others are more skeptical / critical of these possibilities or of what they would entail, while there is also a fair share of disagreement on the timeline of their likely occurrence (if any).

Semi-structured interviews are supposed to be flexible and to elicit meaningful answers. This means that an interview script or guide is prepared beforehand but that according to the development of the interview other questions are also possible, according to the specific answers given by the interviewee and the importance they have for the topics at hand. At the same time, the degree of epistemic trust attributed to these opinions of the experts is not the same it would be if the interviews were to be conducted (or a questionnaire to be distributed) to lay people. Rather, these are knowledgeable inputs that beg a hermeneutic interpretation and a critical discussion – and this is what has been done in this thesis. Moreover, in these interviews I do not refrain from putting forward my own perspective, and the goal was never to reach a consensus but rather to explore the several viewpoints that can be taken on these issues, thus shedding light on how these fundamental issues are looked at by experts at this moment in time. Interview types are also often distinguished according to whether they are “open” or “closed” (Ketele and Roegiers, 1999, p. 21) and according to this typology these were open interviews, insofar as I had no preconceived hypotheses or claims that I wanted to validate. On the contrary, this thesis being an exploratory work, the goal of the interviews was to gather the experts’ insight in order to then assess them critically.

Furthermore, these interviews also add value to this thesis’ scope of communication of science divulgation. The semi-structured interview guide is presented below in Table 06 and the full transcriptions of the interviewees can be found in the Appendix.

Interviews elicit narratives and can be described as some sort of dialogical creation of meaning – they shed light on the phenomena analyzed in the concrete dialogue that is taking place between the interviewer and the interviewee. Indeed, when people tell stories, they select details of their experience from their stream of consciousness. It is this process of selecting constitutive details of experience, reflecting on them, giving them order, and thereby making sense of them that makes telling stories a meaning-making experience. A number of authors were interviewed (11), to gather their different and sometimes diverging experiences and viewpoints, which enrich the debate for an insightful analysis.

In qualitative approaches some distortion can occur because of the role of the interviewer. It is inherent in the interaction between the interviewer and interviewee: The interview is a conversation, the art of asking questions and listening. It is not a neutral tool, for at least two people create the reality of the interview situation.” (Denzin and Lincoln, 2018, p. 900) This was minimized as to ensure validity and reliability whether by giving space to the interviewee to express himself in the course of his intervention whether by avoiding misleading him to conclusions.

Several criteria were taken into account when designing the questions, such as the relevance of the subjects in relation not only to the main themes of this thesis, but also a secondary goal: the interviews sought to elicit some degree of controversy, that is, to push the interviewees to explicate their opinions on the ‘hot topics’ that can sometimes be divisive. Methodologically, and consistently with the format of semi-structured interviews, a number of open-ended questions were designed, which unlike leading questions establish the territory to be explored while allowing the interviewee some freedom. The questions outlined were objectively linked to the issues investigated throughout the thesis and so the answers of the experts, or the conclusions that can be drawn from their answers can therefore also be found in different sections in order to support or refute specific arguments.

To reiterate: the purpose of these semi-structured interviews was to gather data out of the experience and knowledge of the interviewees in order to elicit a discussion over those issues. The interviewees contacted (listed in [Appendix B](#)) are renowned experts in subject areas of research of this thesis.

The majority of the interviewees are gathered around a few fields of expertise, such as neuroscience, computer science and philosophy as to permit comparisons toward a possible trend or pattern found in the subjects discussed. As such it was possible to select a relatively representative pool that sufficiently reflects the range. This is not an exhaustive sample. Other authors were contacted (see the full list of invitations and answers in [Appendix B](#) and not all of them accepted. But the relevance of these experts in their respective fields and the diversity of their backgrounds and opinions make this a meaningful and well-balanced sample. For a brief presentation of the profile and position of each expert, which justifies their selection for the sample, see [The Appendix C - Interviewees Presentation](#). Since self-selection and randomness are not compatible the so-called purposeful sampling (rational choice) was done in the candidates' selection procedure. Furthermore, some of the experts during email correspondence suggested they might not fit in because of their personal beliefs being contrary to many of the issues approached, which to me was precisely a reason to include them. Hence the expected end result should be a balanced and not biased analysis.

Allow me to explain how the information stemming from the interviews was organized. The data was carefully divided into meaningful excerpts of the transcripts. This was done by following the order of the questions as they appeared in the original script. As such, these excerpts were put together to shape the material into a form in which it can better be displayed. Following the order of the questions allows us to categorize the body of work and allows for a thematic connection. This crafted narrative made out of the authors' words is followed by an analysis I put forward. These excerpts are related, in this analysis, to other parts of the thesis, whenever it seems appropriate to do so. It is also important to make clear that the format chosen to display this information stemming from the interviews tries to adopt a narrative form. This is because "the narrative form of a profile allows the interviewer to transform this learning into telling a story (Mishler, 1986). Telling stories is a compelling way to make sense of interview data. The story is both the participant's and the interviewers". (Seidman, 116).

The narrative form of the analysis has been purposely outlined in a story-telling kind of style as to allow the reader to go through each category, each question, and therefore each theme or subject, as one piece of text as a whole without the need to recur to other texts. Each question is then answered by all participating experts and followed at the end by my personal views. These are ingrained with quotes extracted from the experts'

interviews precisely because each word a participant speaks reflects his or her consciousness (Vygotsky, 1987). The participants' thoughts become embodied in their words. "To substitute the researcher's paraphrasing or a summary of what the participants say for their actual words is to substitute the researcher's consciousness for that of the participant." (Seidman, 2006).

Also, and as a supplement to the analysis of the content of the interviews, and in order to better detect regularities in the speech of the interviewees, I constructed word clouds for each of the interviews (displayed at the beginning of the transcript of each interview's transcripts, see each appendix for the respective word cloud) and also a general word cloud (placed at the beginning of the Appendix section) gathering the most frequently used words for the eleven interviews. This is a simple and intuitive manner of literally showing the importance of each topic or concept for the interviewees, through a direct representation of the frequency of its appearance. Not surprisingly, this general word cloud reveals "mind-uploading" and "superintelligence" as being the most frequently used expressions but also, and significantly, "progress". In the Content Analysis below I address the significance of these answers.

In the following sections of this chapter, first I explain how the objectives of the thesis were connected to the questions that composed its script (the guide) and then I explain in more detail how the whole process with the interviewees was conducted.

2 _ Objectives

The objectives of the interviews are harmonized in accordance with the general objectives outlined for this thesis, i.e., each general objective can be linked to the outcome of one or more questions. Hence, the concluding remarks I made at the end of each question's analysis (Analysis #1 @ Analysis #9) can be directly tied to the various objectives as presented in the following Table 05.

Table 05 – Link between the general objectives of the thesis and the questions and respective analysis posed by the interviews

OBJECTIVES	QUESTIONS / ANALYSIS
To explore the link between the convergence of NBIC technologies and the emergence of human enhancements such as mind uploading and the creation / development of superintelligence.	QUESTION II / ANALYSIS #3 QUESTION V / ANALYSIS #9
To design Thought Experiments that are applicable to the main themes of this thesis, i.e., to help confirm or reject some premises regarding mind uploading and superintelligence.	QUESTION IV / ANALYSIS #7 + ANALYSIS #8
To scrutinize the state-of-the-art of Whole Brain Emulation and thus test its feasibility and estimate its timeline.	QUESTION I / ANALYSIS #1 + ANALYSIS #2
To investigate the outcome and impact raised by Mind Uploading regarding consciousness and personal identity issues, eventually altering what it means to be human.	QUESTION I / ANALYSIS #1 + ANALYSIS #2 QUESTION V / ANALYSIS #9
To critically analyze both the claims of Mind Uploading's proponents and of its skeptics so as to cover it as extensively as possible and thus enable an open debate over open questions.	QUESTION III / ANALYSIS #5 + ANALYSIS #6
To apply the MU and SI thought experiments described in Part I, by analyzing their content in the context of mind uploading and superintelligence, and therefore verify their internal consistency, hence validation.	QUESTION IV / ANALYSIS #7 + ANALYSIS #8
To clarify Mind Uploading's purpose and meaning, thus trying to prove its importance and significance for humankind's evolution.	QUESTION I / ANALYSIS #1
To discuss about a possible clash between an Artificial Superintelligence and our own and argue in favour of Mind Uploading as a technology to help humankind in that competition using the designed thought experiments to support that dispute.	QUESTION IV / ANALYSIS #7 + ANALYSIS #8

3 _ Interview Guide

As I explained in the intro to this chapter above, the main method used to gather evidence in this part of the thesis is the interview. However, even though the semi-structured interview allowed for a great degree of flexibility, I decided to send the script (the guide)

to interviewees beforehand, as if it was a questionnaire. This was meant to prepare the interviewees for the discussion that was to come, so as to allow them to reflect on their positionings even before the interview.

This interview guide comprising the questions listed below (Table 06) was then followed in the interview, though not strictly, which very much depended on the fluency of the interview itself. The end result was successful as we can say that the questions were in general answered, but also that many other meaningful insights stemmed from non-predicted (or predictable) twists and turns of the dialogue that arose through the interviews, and this only enriched the content that I will analyze in section 5 below. The guide that was sent is the following.

This guide was already included in the project of this thesis that was approved at its outset. It was revised by my supervisors and approved by the jury of the project. It is meant as an instrument in the discussion and communication of science, following advice of the coordination of the doctoral programme in which this thesis is enrolled.

No pre-test of the interview guide was done, as might be expected in the cases of interviews conducted with non-experts (the general public) for one simple reason: experts in these fields do not abound, so it would not make sense to test the interview with an expert that would not be suitable for the real interview afterwards. As such, after distributing the guide among interviewees, the interviews were conducted and given the enthusiasm with which I was met by some of the interviewees, they went in different (and sometimes unexpected) directions.

Table 06 – Interview Guide.

Question I _ <i>Regarding Whole Brain Emulation and mind-uploading</i>	
Q1: Would another me be me? (or just a copy of myself)	
If so, then	How would you know you are you?
If not, then.....	What would our copies <i>really</i> be?

Q2: What would represent the biggest risk in achieving something like this?

Question II _ Regarding Superintelligence and the Technological Singularity

Q1: If (*When*) it happens (*and we're still humans*) will we even know it (*be aware of*)?

If so, then Could it be because we would be *less bio and more bit* at this specific moment in time?

If not, then..... Would it be because we would have failed to cope with progress?

Q2: How possible would it be for a *Singleton* of some kind to rule the World?

Question III _ Regarding Transhumanism and Posthumanism

Q1: Will *being human* be different in a Post-human Era than it is today?

If so, then Would this mean we would definitely be *more bit than bio*?

If not, then..... Would this mean that an eventual cyber-sapiens of some kind would in fact be the next natural step in human's evolution?

Q2: Do you think politics in a few years will be divided between bioconservatives vs transhumanists?

Question IV _ Regarding Superintelligence and mind-uploading

Q1: Would it make sense to iterate human intelligence via mind-uploading to augment it / improve it, and in this manner cope with AI's progress?

If so, then..... Could we avoid a singleton of some kind or AI's progress would sooner or later always develop into a <i>Superintelligence</i> ?
If not, then..... Would <i>universal darwinism</i> not select AGI as the <i>more adapted intelligence</i> instead of ours?
Q2: Could mind uploading prove that two different persons sharing an identical blueprint would in fact be identical? And if so, would this not shake our beliefs in issues such as identity and personality?
Question V _ Regarding Transcending Biology and Human Enhancements
Q1: Is it inevitably going to generate a dystopian society?
If so, then Could/Should we trust our fate in a system run by an AGI?

4 _ SSI - Rationale

Before we enter into the analysis of content properly speaking some final notes on the concrete steps taken for the interview process are due. As already mentioned, in order to adequately address this thesis' objectives as shown above on Table 05, and after an analysis of the technologies linked to human enhancement, a first step was the constitution of the sample, through a selection of the experts on these matters (sampling by rational choice) which was conducted through a previous literature review based on the papers, books and research projects of the experts important for the fields of the topics analyzed in this thesis. [The Appendix C - Interviewees Presentation](#) exhibits the kind of profile of the authors interviewed for this thesis and the full transcriptions of the interviews can be found immediately afterwards in that Appendix section by a chronological order of interview (all held between March and April of the current year 2020). The second step taken was sending out the invitations A chart listing every contacted expert/author (including the outcome of the invitation, whether this was a positive or negative response, and also the cases in which no response was attained)

can be found in [Appendix B](#). I invited a total of 37 experts and of these almost a third (11 interviews) were successfully held. One limitation that might be mentioned, and that probably impacted the number of interviews done, was that the contacts and the interviews were done at a time when the Covid-19 pandemic hit and many places of the world were going into lockdowns. This, alongside with the sheer status of the interviewees (and the fact that many of them have very busy professional lives and as such are hard to get a hold of) perhaps explains why some of them declined to be interviewed or did not answer. Even so and taking into account that these experts were being contacted by an “unknown” Portuguese PhD candidate with no previous reputation in the field, I consider that having been able to conduct as much as eleven interviews with these reputed experts can be considered a good result. A third step, taken as the invitations were sent, was the attainment of an informed consent (see [Appendix A](#)). Also, and just before the interviews started, recording of the interview as well as the discussions’ transcriptions was asked and approved by the interviewees to ensure no proprietary breaches of confidentiality. The interviewees were assured that their interviews would be used only for the academic purposes of this thesis and any further use of them will of course require their explicit consent.

The interviews were performed through videoconference (Skype) but also through e-mail (Chalmers and Cattell) always depending on the availability of each expert. It goes without saying that the semi-structured interview only fully achieves its goals when a live interaction takes place, and so this grants the interviews conducted by e-mail a different status than those that were done by skype. We can note, for instance, that the answers in the interview with Chalmers are much less developed than the majority of the answers attained in the other interviews. Again, I respected the decision of each interviewee according to their own availability, and we can say that the overwhelming majority of interviews provided a very rich and often carefully argued material to be discussed in this thesis and to serve our purposes of communication of science. The overall goal was to understand and explain in what way the experts within these domains of *NBIC-technologies* are concerned with human enhancement and its outcomes, focused on the researcher’s specific ideas, and how they can be related to the dissertation thesis of *Whole Brain Emulation, Mind-Uploading and Superintelligence* and its specific objectives.

This semi-structured interview for conducting qualitative research was designed to incorporate theoretically driven questions eliciting data grounded in the experience of the

interviewee and to get in-depth information on topics of interest discussed throughout the thesis. Thus, interviewing distinguished authors was chosen for the current project because it provides an opportunity to generate rich data and because they represent the cuttingedge experts in these research areas.

The following Table 07 and Table 08 show on one hand (and on the left side of each table) the summary of the main questions followed by their additional questions made to all the authors (interviewees). On the other hand (and on the right side of the tables), a summary listing the interviewees, the type of interview held, the time it took and the main field of interest of the author is exhibited. These tables reflect an overview of details regarding the Interview's Rationale altogether.

Table 07 – Overview and summary of questions and interviewees with their fields of expertise.

FIELD OF INTEREST	MAIN QUESTION	ADDITIONAL QUESTION	INTERVIEWEE	TYPE / LENGHT	MAIN FIELD of INTEREST
WBE / Mind-Uploading	Would another me be me? (or <i>just a copy</i>)	What would represent the biggest risk in achieving something like this?	RANDAL KOENE	Skype / 50min	NEURO_ SCIENCE
Superintelligence / Technological Singularity	If (When) it happens (<i>and we're still humans</i>) will we even know it?	Would it be possible for a Singleton of some kind to rule the World?	DAVID CHALMERS	e-mail	PHILOSOPHY
Humanism. Transhumanism. Posthumanism.	Will <i>being human</i> be different in a Post-human Era than it is today?	Do you think politics in a few years will be divided between bioconservatives & transhumanists?	ANDERS SANDBERG	Skype / 60min	NEURO_ SCIENCE
Superintelligence / Mind-Uploading	Thought Experiment HUMAN PREDICTOR RESPONSE	Thought Experiment ITERATING HUMAN INTELLIGENCE	KEN HAYWORTH	Skype / 65min	NEURO_ SCIENCE
Transcending Biology/ Human Enhancements	Is it inevitably going to generate a dystopian society	Until what extent are enhancements ultimately leading life extension?	NICHOLAS AGAR	Skype / 55min	PHILOSOPHY
			Expert / Author	Skype / e-Mail / Phone / in Person	15 min 30 min 45 min 60 min

Table 08 – Overview and summary of questions and interviewees with their fields of expertise.

FIELD OF INTEREST	MAIN QUESTION	ADDITIONAL QUESTION	INTERVIEWEE	TYPE / LENGHT	FIELD
WBE / Mind-Uploading	Would another me be me? (or <i>just a copy</i>)	What would represent the biggest risk in achieving something like this?	JAMES BARRAT	Skype / 50min	FILM_ DIRECTING
Superintelligence / Technological Singularity	If (When) it happens (<i>and we're still humans</i>) will we even know it?	Would it be possible for a Singleton of some kind to rule the World?	RICK CATTELL	E-mail & Skype / 30min	COMPUTER SCIENCE
Humanism. Transhumanism. Posthumanism.	Will <i>being human</i> be different in a Post-human Era than it is today?	Do you think politics in a few years will be divided between bioconservatives & transhumanists?	ROBIN HANSON	Skype / 45min	ECONOMY
Superintelligence / Mind-Uploading	Thought experiment HUMAN PREDICTOR RESPONSE	Thought experiment ITERATING HUMAN INTELLIGENCE	TOM KOCH	Skype / 60min	MEDICAL ETHICS
Transcending Biology/ Human Enhancements	Is it inevitably going to generate a dystopian society	Until what extent are enhancements ultimately leading life extension?	DAVID PEARCE	Skype / 35min	PHILOSOPHY
			RICHARD LOOSEMORE	Skype / 75min	COMPUTER SCIENCE

5 _ Content Analysis

As mentioned in the introduction of this chapter, the content analysis is made up of selected transcript excerpts of the interviewees (*by order of interview*) regarding each question followed by my concluding remarks – Analysis#1 @ Analysis#9. Questions are highlighted in light-grey to stand out and serve as a reference in the content analysis and are inserted together with the excerpts in tables only to easily distinguish them from the analysis made at the end of each section.

Table 09 – Selection of transcript excerpts regarding Question I – Q1.

Question I
Regarding Whole Brain Emulation and Mind-Uploading
<i>Q1: Would another me be me? (or just a copy of myself?)</i>
RANDAL KOENE: What if we used a MU-procedure in which you don't copy yourself to some other thing that is far away but instead you're replacing everything that is neurons and so you're putting artificial neurons in your head until you have an artificial brain, at no point was there ever a very clear distinction as to whether you switched over to a completely artificial brain so it feels like it still is me. If you were to replace the neurons at a rate where you would actually get it done in your lifetime you would have to be replacing millions of neurons per second so even that gradual approach would still be something that would feel a lot like the whole copy scenario." "Those of us who are functionalists, tend to feel that consciousness is emergent from the brain, so that shouldn't be a problem. You can generate consciousness in the artificial brain just as much as you can in your own brain. Personal identity is how all of what we're experiencing right now gets processed through our memories through all of the synapses that have been changed over time since we were born, but even

the settings of the synapses based upon whatever the blueprint was when we were born.”

“The thing that makes it unique that makes it me, that is my identity, is the colour that I give to it. The experience that I go through because of all of the context that I have because of what it means to me. Your identity is encapsulated in your memory, not in your conscious momentary experience. Your pattern, that identity is still mostly there, if you can reconstruct those memories. And I say mostly there because I think the whole concept of identity being this sort of uniform this one thing that you can point to and say this is always me, is already a case of, I guess it’s maybe not an illusion but it’s kind of a wrong way to talk about it that we tend to be used to. So, there is this aspect of identity that is this changing thing and now the thing that gets added to it, when you talk about something like MU is the possibility of having points in time where not only are you going from a previous to a next identity, you may also be branching! So it could be, if someone comes up with a MU technique that preserves the original brain in the process, we could end up with an artificial brain and a biological brain, and at the moment of creation - this is both you! But as soon as they start to experience slightly different things, they begin to become different, so you have a branching going on.”

“So, I think there is a visceral reaction that sometimes goes back to a real personal philosophical difference and sometimes it’s just visceral and just emotional. Just something we haven’t really thought it through even if we actually do share the same philosophy.

ANDERS SANDBERG: “The issue here is of course of personal identity as well as a related question: can you have a conscious mind existing in a software? In a functionalist perspective on philosophy of minds you could probably have software minds. I also don’t think personal identity is a problem because I don’t think personal identity exists. I actually use Derek Parfit as my landlord.”

“When we get to brain emulation, if it really goes very smoothly where we do it continuously, it prompts our intuition to say yes that’s going to be a continuation of ourselves, our narrative is not broken. I think it is more likely for at least early brain emulation, there is going to be some dramatic shifts here in how you exist. First you are a living person then you undergo this process and then you are a software person.”

“It doesn’t disrupt things too much and indeed I’m very happy with multiple instantiations of me. We’re all Anders.”

DAVID CHALMERS: “They’d be two different people. Both might have equal claims to being identical to the prior person.”

KENNETH HAYWORTH: “There is no me. The idea of a unique soul being part of an individual is a primitive concept that really does not stand up to our current understanding of neuroscience and the brain. There are forms of computation of the brain where consciousness comes from and our sense of the brain and our sense of identity comes from episodic memories that we refer to. A Whole Brain Emulation would be you in that sense.”

NICHOLAS AGAR: “I think that there’s a difference between copying and a set of processes that suffice to transmit identity or survival. You could copy me precisely, but I shouldn’t view that copy as a continuer of me. I disagree with views like the psychological continuity type of personal identity.”

TOM KOCH: “If they’d put it in a body of a twenty year old they would put an old man into a young man’s body, and the two personalities would differ. And in that environment would I think of myself as a material being or would I be just a spectre of a disembodied thought?”

“Science Fiction has two models here. One in which you think that you’re in fact a whole person and the other one in which you know that you’re a cyber construct, basically a ghost of yourself looking for a way to inter-react with the world that you have left. I don’t really believe that these are anything other than purposely philosophical questions. I think that the idea of the singularity is one of those thought experiments which is taking out reality.”

JAMES BARRAT: “I think that you would be just a copy of yourself. I don’t know how you would transfer the essence of someone to another me. Just because we don’t even know what we’re talking about when we say things like soul or essence or our consciousness.”

“To me the whole mind uploading is about living forever. Transhumanists, like everybody else, have a deep fear of dying. This is like one of those created religions with religious promises that you will never die. I became skeptical, and not just of the ability to do it but the motives behind the people who want to do it. We could end up making just empty vessels. Empty emulations.”

RICK CATTELL: “The biggest problem with emulation would be the loss of my biological body and everything that comes with it. However, it could be an arbitrarily good emulation, such that people would consider it a way to “survive” dying... “

“If you’re talking about my mind uploaded to a virtual world that is difficult to distinguish from the real world, then the biggest problem would be emulating the real world in a way that I could not distinguish it from the one I knew. And if I knew it to be a virtual world (asferent they discover in the movie Matrix) it would be a psychological problem, exactly because it’s not real. Psychologically I want to feel like I’m real.”

ROBIN HANSON: “That’ll be relatively context dependent in the sense that different people in different cultures and different ways that they were brought up will think differently.

DAVID PEARCE: “I don’t think classical digital computers can solve the binding problem. Essentially, I think what defines our minds is local binding by individual perception of objects, and also global binding a unitary self. I don’t think classical digital computers can emulate this even if one were to replace the ones and zeros of classical Turing machines notionally with discrete pixels of experience. You will have a micro-experiential zombie. Phenomenal binding is classical impossible. It is possible to use digital computers for a backup-system but I don’t think that what we backup is a subject of experience.”

“I think identity is a convenient myth, but how much this matters is debatable when one wakes up in the morning, one isn’t really troubled “Am I the same person who fell asleep last night?”, and I think some people would regard mind uploading as akin to the dilemma one has waking up in the morning but in practice I think it’s much more radical than that, because essentially it’s creating a zombie.”

RICHARD LOOSEMORE: “Everything depends upon the definition of me. In our entire experience the concept of me never involves any possibility of a duplicate. As soon as you stretch to situations in which it either gets duplicated or transported, you’re already outside the bounds of what the whole concept of me was meant to address. Strictly speaking the question can’t be asked because there is no concept of genuine versus fake inside the original concept because it never happened. So in a sense they both are you.”

“If there’s a WBE of me, that is complete and full, and if I expire, if I die in the process of creating, if I destructively upload myself then it seems easy to say that that new me is the same as me. It’s not correct, it’s a new me, but on the other hand it’s not any different from any other situation where you go to sleep and in the morning the whole

system comes back up again and we never ask the question is the me of this morning the same that went to bed last night, because there's a physical continuity. But the physical continuity doesn't matter an awful lot. There was an interruption in the consciousness stream and then a resumption and the fact that it's interrupted and resumed every single night is not amazingly different from an interruption that also involves physical movement to another substrate."

Analysis #1:

For Koene our identity is encapsulated in our memory, not in our conscious momentary experience. Our pattern (of information), that identity, is still mostly there in an upload if one can reconstruct those abovementioned memories. But as soon as it starts to experience different things it becomes different, so there's a branching going on. For Sandberg (who is also a functionalist) we could probably have software minds (uploads). He doesn't think personal identity exists and has no problem with multiple instantiations of himself. They all would be Anders (Sandberg). For Chalmers the copies, or the copy and the original, would be two different people. For Hayworth there isn't me. Our sense of identity comes from episodic memories and a Whole Brain Emulation would be you in that sense.

The notion of a "pattern of information" is a term used by many authors (e.g., Moravec, 1988; Goertzel, 2006) to express the also ambiguous concept of the mind. It is commonly used in reference to a set of memories and patterns of thinking, and many of them believe that identity is ultimately a question of preserving informational patterns; hence there is a general adoption of this patternist view for what we usually call the mind and personal identity. (Wiley, 2014; Cerullo, 2013; Bamford, 2017). Ray Kurzweil (not interviewed for this thesis) even compares identity to this pattern: "my identity is rather like the pattern that water makes when rushing around a rock in a stream. The pattern remains relatively unchanged for hours, even years, while the actual material constituting the pattern of the water is replaced in milliseconds." (Kurzweil, 2000). Koene identifies (elsewhere) this pattern as the information content of our minds: "By the information content, I mean both the parameter settings (e.g., memory), as well as the ways in which

the parameters are used, the functions carried out by the mind... That pattern is all that we are aware of being.” (Koene, 2011). Again, in the interview for this thesis, and specifically to answer the question above, he underlines this patternist concept.

Agar doesn't view that copy as a continuation of himself. Koch questions if he'd be just a spectre of a disembodied thought. Barrat thinks an upload would be just a copy and fears we could end up making just empty vessels. Empty emulations. For Cattell an upload in a virtual world would be a psychological problem since psychologically one wants to feel like one's real.

Hanson says it's context dependent in the sense that different people think differently. Pearce thinks that our minds are defined by local binding by individual perception of objects, and also global binding a unitary self. He doesn't think classical digital computers can emulate this and he doesn't think that what we'd backup would be a subject of experience. He thinks identity is a convenient myth and some people would regard mind uploading as akin to the dilemma one has waking up.

Loosemore also thinks everything depends on “the definition of me”. In our entire experience the concept of me never involves any possibility of a duplicate. In a sense both of them (the duplicates) are me. He makes reference to the similarity in going to sleep because of the physical continuity. But he stresses out though, that an interruption in the stream of consciousness and then resumption would not be that different from an interruption that also involves physical movement to another substrate, which would then mean we could theoretically copy ourselves to another substrate and it would still be us, at least for that moment on.

The first thing that strikes me in all the interviewees replies concerns the way they position themselves as being for or against it – with “it” being the singularity and/or what comes along with it, such as possible whole brain emulation and mind uploading. These (WBE and MU) are taken as side-effects of a quite controversial theme which prompts some of the interviewees to think that these possibilities simply do not make sense because they are never going to happen. As such, it would seem that some of these views are conditioned by some sort of personal philosophical standpoint. If an analogy is allowed here, it is almost as if a question such as “did a big bang ever occur” quickly would have divided religious people from non-religious ones, insofar as a deeply ingrained belief seems to be dividing two camps in these issues.

We can find another division among the interviewees relating to the difference between the scientific view and the more philosophical one. Neuroscientists (mainly functionalists) generally tend to accept the possibility of whole brain emulation and mind uploading but refer to personal identity as something that exists essentially in our own and personal minds. Something that maybe has a psychological continuity but also something that does not really exist out of the context of our own choices and personal philosophical beliefs. Therefore, they can be emulated either because there wasn't a self in the first place or because there may be two selves in the second place (after WBE).

The point in which most interviewees agree is in the fact that whatever it is that defines me as being me depends on time and experience. As soon as time goes by, the new me starts its own experience and branches off the old me. So, the general conclusion out of these answers to this specific question would be that, if copying myself through mind uploading / whole brain emulation would become possible, it would always generate a new experiencing self with, from that point on, my memories as a basic infrastructure of that experience. But from there on, the new self, would definitely become a different person.

The transfer of personal identity is one of the issues detractors of MU usually argue with to question its successful outcome. However, and as these interviewees' answers reflect (as discussed in the MU chapter, Part III) identity theories are far from consensual. Moreover, it does not seem unreasonable to infer that whichever the outcome of MU will be, MU itself may prove to be decisive in establishing new identity concepts.

Table 10 – Selection of transcript excerpts regarding Question I – Q2.

Question I
Regarding Whole Brain Emulation and Mind-Uploading
<i>Q2: What would represent the biggest risk in achieving WBE/MU?</i>
RANDAL KOENE: “Some of that risk depends a little bit on what a person or a society considers good versus bad, whatever we assign those criteria to. Many of us would find it bad if uploading were made possible and then you end up in a world where there are only 5 million copies of the same person and no others? Suddenly you have very little variation, differentiation, all of the other branches have basically ended and there’s only just the one that proliferates.”
“There would be a risk involved if having MU available leads to either the eventual disappearance of regular biological homo sapiens or if it leads to a situation where, because there are ten trillion uploads and only seven billion biological homo sapiens that in terms of the amount of influence that one or the other set of people have that it’s huge to have both and the same could be said what if the mind uploads can all process things ten times faster and have other advances (e.g. mental advances), and develop mental possibilities that give them an edge? This could be perceived as a risk towards original biological humans.”
“On the flip side though, the whole point of going through MU is really the advancement of the species. I find it really hard for example to imagine that you would ever encounter a space-faring species that goes from star to star that is embodied as a sack of blood and tissue that needs to be oxygenated than something that is more suitable for the environment of space. The species that is advanced would want to augment and adapt themselves to different circumstances, because things don’t always stay the same. There will be competition between artificial general intelligence and humans and the challenges will be different. We need to be able to direct our own evolution into the directions where our chances are still being a part of what goes forward and taking advantage of possibilities and developments where our chances are simply better.

Those would be the ones that have more adaptability and that have evolved themselves further. And this (Mind Uploading) is a kind of super adaptability. We're beginning to see intelligent agents that come out of AI that will become more competitive. So, you end up having an ecosystem of intelligences and we fit in there somewhere. So, the question is what role do we play in that ecosystem of intelligences, and that's why this kind of development is very important."

ANDERS SANDBERG: "There is some fundamental scientific and philosophical or technical reason why it cannot be done: Is there enough scale separation in the brain to make it possible to emulate a brain properly? If you don't simulate it on all the scales, you get what computational fluid dynamics people call dry water. It's like water but it's not water."

"I do believe that we do get a fabric once we get down to the subcellular model following the standard computational neuroscientists assumption: You can divide the neurons to isolate compartments and that's well enough. Inside, things are mixed because the Brownian motion is effective over those distances. At some point you have scale separation. But this remains to be proven, and if it's not true, in that case you might need to simulate things on a finer degree. So, we'll eventually run into hardware computational problems, then it might be a problem in principle to copy things."

"Then we have the philosophical issues. Philosophy of mind is hard, and I don't think we understand consciousness so we might find out something surprising."

"I can take my super-computer and run a very nice set of Hodgkin-Huxley equations for a big piece of simulated cortex and get roughly a course of activity. That's not the same thing as WBE. We've done small models of the cortex. The Henry Markram Blue Brain Project demonstrates beautiful stuff that looks biological, but we can't test it yet. We're moving in that direction, but we actually haven't got the right proof yet. It's going to be a very tough project to actually make a working WBE. But of course, many other people would say wow you think that there's even a chance that it happens that is mind-blowing. So, this is why I have this kind of position in the middle."

KENNETH HAYWORTH: “I think that the biggest risk is that it might work! And you might end up in a situation that is bad.”

“My neuroscience colleagues, if I open up the time-frame as wide as possible, they’ll say yes. If I say, given what we know and our current methods of preservation of sufficient memories eventually they will say yes this seems to cover all the basics.”

NICHOLAS AGAR: “a digital copy of me that maybe starts out sharing my interests and wanting to preserve me, but over time develops quite good interests and notices and says sort of well, this earlier human being from whom I was copied is quite susceptible to this coronaviruses, I’m not susceptible, and do I really care? A little bit, but not much. As soon as you do this (Mind Uploading) you do almost inevitably create a being that has certain interests. And I don’t know if I would want to.”

“So, my children for example. I guess I care more about their interests (they are flesh and blood), than I would do right now about a digital me. And I wouldn’t want a digital me to say well, screw your dumm children!”

“When you have a technological ideal, it’s incumbent on you, it’s important that you would ask what happens if we would fall short? Any technology is worth pursuing, but you should be aware of near misses. And achieving that ideal can be pretty disastrous. So I would ask the same general questions about uploading as soon as we (if we) would succeed. Then you’d have a lot of possibly considerable morally beings...”

TOM KOCH: “In either situation the risk would be that the digital version finds itself to the extent that it knows what it has lost. There is a long tradition that argues that mortality, that the very nature of our existence has more to it in finite beings. Much of that contributes to the value we place in life and on the immediacy of the day. Would have Albert Einstein been Albi if he wouldn’t be able to play his violin? The violin was in many ways a signature of a kind of thinking that he brought to his work. One of the deep questions in proposing this (Mind Uploading) as a solution to what may not, be a problem but a fear, we in fact attack one of the very Essence of the things we’re used to give life meaning. And frankly, how many of the really frivolous rich do we want to have preserved. The final problem is really who is going to be preserved? How we

can set up those criteria of who will get it will always be questionable.”

JAMES BARRAT: “As I said in my book *Our Final Invention*, the people who will get augmented or who will get the brain enhancement won’t necessarily be the most ethical. The other big danger, besides letting psychopaths take over, is if everybody lives forever what about the new people being born?”

“And this is a side-effect another implication of brain augmentation and brain uploading, but if you have a bunch of people living forever, you’re by necessity excluding a whole other bunch of people from living meaningful lives that are running out of resources. Look at the carbon footprint of our artificial network. It’s astronomical. Cyberspace does not take zero resources. And these people would be consuming resources and demanding services and doing other kinds of things rich people do.”

RICK CATTELL: “Ultimately, I think the biggest risk is that we are all living on one small planet, so that a mistake could lead to our destruction as a species: First of all, there will eventually be exhaustion of resources: that is, if people can propagate by self-duplication, we will eventually run out of real-world resources to sustain ourselves. Second, there is a problem with our evolution. Improvements in both individuals and societies have been founded on competition that has ultimately been made possible by the death of existing individuals and societies. Third, we will not avoid the temptation to make improvements to ourselves as we duplicate ourselves, which may backfire if we take a step in the wrong direction. It’s an intriguing topic. It’s probably a few decades too soon to have any real impact on us, but it’s good to think about the philosophical morality and the possibilities...”

ROBIN HANSON: “Well as always the biggest risk is extermination or extinction. And not quite extinction but a large disruption that people would not agree in the new relationships where who’s in charge, or who’s more dominant, more rich, more respected, because these sort of changes, change those things.”

DAVID PEARCE: “Let’s assume that life has to be cherished, protected and preserved, then mass destructive mind uploading would in theory be a form of existential risk, inasmuch as one would be creating effectively zombies.”

RICHARD LOOSEMORE: “It does depend upon how you do the emulation. I tend to believe that you actually don’t need a neuron to neuron let alone a molecule by molecule reconstruction of the brain. You can get the entire software emulated very thoroughly down to the level, so accurate that it’s down to the level of little bits of mind or noise that you really don’t even care about as a personality without having to really map every single neuron. If there was a way of figuring out the functional state and a functional diagram, sort of functional architecture well enough that we know how it works and why all the physical neurons play the role they do, it might turn out that, a cluster of a thousand neurons play the role that can be replaced by one little routine. And it’s exactly the equivalent that plays the software routine we know why and we’re sure of it so therefore the upload doesn’t have to involve an absolute copy.”

Analysis #2:

Beyond the obvious existential risk pointed out by each and every interviewee, as expected, and the other obvious fact that we are dealing with something quite dangerous with WBE & MU, the interviews also reveal some innovative views.

R. Koene admits that MU could end up generating millions of copies of this one person in detriment of all others leading to a lack in variability and differentiation and also that uploads could process faster and be mentally more advanced, which in turn again could lead to an eventual disappearance of the regular biological Homo Sapiens. On the other hand, Koene argues precisely this to be the motive for developing MU, for it stands as an edge our species may have in augmenting capacities and hence competitiveness, in an ecology of different intelligences as may be the one ingrained of AI clusters in the near future. Hayworth sustains that given current methods of preservation of sufficient memories a WBE would actually be feasible and claims that his neuroscientist colleagues generally tend to validate this as long as in a wide-open

timeframe whereas Sandberg seems more skeptical of its achievement due to technical issues. He wonders whether there is enough scale separation in the brain because simulation has to happen on every scale, to make it possible to emulate a brain properly. Because of this he does predict eventual hardware computational problems ahead in this grand project of WBE. Loosemore on the other hand thinks that we won't actually need a neuron to neuron let alone a molecule by molecule reconstruction of the brain, as long as we'd figure out the functional architecture, that we knew how it works and why all the physical neurons play the role they do. It might turn out that a cluster of a thousand neurons play the role that could be replaced by a subroutine.

So, there is this evolutionary approach arguing in favor of mind uploading as an incrementing adaptability we would achieve as a species, whilst on the other side interviewees tends to stress the ethical and philosophical issues but also sustainability reasons. Agar fears that MU copies of ourselves would eventually develop their own different interests which are precisely the reason for his worries, since we wouldn't know what kind of interests these would be, and in taking such a technology forward we'd always have to be careful with possible near-misses. Barrat and Cattell underline the danger in eventually running out of real-world resources for sustaining these Mind Uploads. Barrat and Koch also fear the criteria for getting a WBE since they suspect that the people who would get augmented or brain enhanced wouldn't necessarily be the most ethical. Koch also argues that our existence has more to it than the goal of MU and a possible immortality of ourselves through it. He recalls the traditional argument according to which finitude gives life a meaning. But he also points out the danger in an upload becoming aware of what it has lost (which by the way, implicitly attributes consciousness). Hanson makes a reference to the sociological risk of having all our relationships changed through MU. Pearce even talks about the possibility of mass destructive mind uploading being a form of existential risk, insofar as one would be effectively creating zombies.

Interestingly, but not unexpectedly, the interviewees identify a multitude of risks and dangers around MU & WBE, meaning that the minimum we can expect from this possibility is that it would completely change life as we know it. Moreover, in societal terms its impact is potentially huge, which prompts us to ask: Why are we even pursuing this? Is it not going to affect humankind in all its boundaries? Is it not going to endanger our very existence? Yet it is being pursued, and as thoroughly discussed in the WBE and

MU chapters (Parts II&III) of this thesis, investing on it and targetting the scientific and technological endeavour it would amount to is only comparable to Man's landing on the Moon.

On the other hand, something that is argued by many as an adaptive technology or means to improve our ability, and even chances to survive (in an ecology of intelligences, with competitive AIs, or in space travel), is the very same thing that basically could lead us to extinction. But if we take a look at the lessons history can teach us, we find an important argument about disruptive technologies and the effects they had on humankind and on its societies. Generally speaking, we cannot escape the fact that technologies have almost always imposed themselves whether for good or bad uses. One conclusion might be that as long as ethical guidelines allow it, the technical means to achieve it will definitely be carried out. The competitive edge that MU might render humankind can actually ultimately be regarded as a means to avoid extinction as discussed in chapter MU (Part III), and particularly in chapter SI (Part IV), which to some extent ironically, could make MU not an existential risk but rather a survival insurance.

Also, the fact that these uploads would probably increase sustainability worries, that are even now (2020) becoming an ever more serious question with the incommensurable misuse of resources we have been doing, is a thoughtful argument some of the interviewees made. However, it is not less speculative, since we might also argue that these uploads could actually reverse that sustainability unbalance in their favor. It could turn out that they might, as envisaged by many authors, do exactly the contrary, thus arguing in a speculative way over a speculation itself can always go either way. Nevertheless, it can be in fact, another headache for MU proponents to worry about.

The selection of uploads (or rather - for an upload eligibility) was another controversial issue raised by some authors, tied to an eventual risk. Who gets to be uploaded? As discussed by some of the interviewees, probably the most unethical (powerful) people would be the ones augmented via MU. This of course might happen, but as I stressed out in the interviews, I think it basically sums up to what we call democracy inasmuch as, like anything else, the political system we are tied to will ultimately be held responsible for making that call. It finally leads to another question following up ahead in these interviews and is also addressed again in the MU chapter of this thesis.

Table 11 – Selection of transcript excerpts regarding Question II – Q1.

Question II
Regarding Superintelligence and the Technological Singularity
<i>Q1: If (When) it happens (and we're still humans) will we even know it (be aware of)? Could this mean we'd be some kind of a cyber sapiens by this time?</i>
RANDAL KOENE: “If you have a superintelligence that has its own agency that has its own goals, and that SuperIntelligence wants to create the sort of technology that will allow it to pursue those goals in the universe, that it will need to put together some kind of resources that we currently have available on earth. I think that it is more likely that we would notice it than that we wouldn't notice it.”
“Also, the AI is bound to the laws of physics and therefore every advance that it makes, every improvement that makes in its process or whatever, requires very sophisticated factory capabilities, and we would notice if things were being produced on this planet.”
ANDERS SANDBERG: “Basically what is the speed of take-off. The way of measuring it is to think about it relative to human reactions. So, you could have a hard take-off but it's false to compare to human reaction times. Or you could have a slow one where our society and the institutions can change at the pace when it is happening. And then you can imagine some middle ground where it's happening rapidly but it's still time where you can make some decisions. I don't think we have actually any good evidence for whether any of these three is the most likely.”
“There are interesting arguments like Robin Hanson's economic arguments pointing out a hard-technological singularity take-off but it's a kind of very weird claim about economics, saying that a subsystem can become the whole and grow much much faster than any other part of the economy. That is a weird claim to an economist. You could make the same argument in ecology or biology and say could a small part of the biosphere take-off suddenly and change geography? And at this point, the biologist

might look around at humans and say oh actually it seems to be happening right now. Because from an evolutionary standpoint what has been going on is, up until recently the biosphere was kind of evolving at a normal pace and then some species goes around the whole thing and changes the rules for evolution. That is kind of an evidence, but you can't really make something relatively fast. Unfortunately, neither of these tells us something about time-scale."

"It's probably not going to be superfast. There might be problems about gathering evidence and data to be effective at being intelligent. But then again it might be much faster than human reactions."

"Another issue of course is how visible this is? We could also say now we have a piece of software that is outrageously intelligent, and we don't even notice how things are going on just as mice can't notice the clever plans humans are making around them. I don't think we have any good tools to analyze this. We can analyze the different scenarios and try to see if we can do something sensible about this and then try to take the right kind of action. Eric Drexler has a critique about the AGI idea in these comprehensive AI services and points out that actually AGI is probably not what most of the companies would like to develop. The business case for AGI is very weak. But you could imagine something like the Amazon Cloud for intelligence. It should be useful if you could connect various sources of intelligence and skills to various tasks. So in this very big wide paper he sketches out a pathway of the direction that seems to be very much likely then in that scenario you don't end up with a technological singularity in the sense of one little server somewhere waking up and taking over but rather that suddenly you get intelligence on tap. We get as much intelligence in the world as we connect data servers to the cloud. And of course, we can use these servers to come up with better ways of making better servers and better technology. So you get a formal augmentation here which might very well mean that some people would augment themselves etc rapidly in an extremely strange situation. But it's not based on any particular single superintelligence existing, it's rather superintelligence as a service."

KENNETH HAYWORTH: "I think we would be aware of it because I see this more as a gradual thing. The AI in its current state had great leaps but it completely fails at

things that no human being would fail at. Essentially no common sense. And I think that what we will see is more of a gradual progression where we will be extremely aware of.”

—

DAVID CHALMERS: “Yes, it will be obvious.”

—

NICHOLAS AGAR: “Nick Bostrom’s book on superintelligence, it seems to me that if I had a complaint about it, it would be that it has too many arguments in it. Well I’m more scared of climate change from an immediate extinction.”

—

TOM KOCH: “All of this talk about superintelligence never deals with an abstract of intelligence without an emotional and an historical context would to me be no intelligence at all. Machines don’t learn. Machines are trained through very precise algorithms to very, very specific tasks. It’s not deep learning! It’s deep training. What you call Artificial Intelligence I would call basically a narrowly trained machine that can replicate the basics of the form and can give us something very likable and advanced. So I don’t think this is intelligence at all, I think they’re basically just better at copying.”

—

JAMES BARRAT: “Stephen Hawking said, superintelligence would be able to create weapons we wouldn’t even understand. In fact, we wouldn’t even understand anything it did. If somebody or something is like a thousand or a million times more intelligent than we are, in a mathematical rational sense, we really wouldn’t understand anything about it. It would be like suddenly sharing the planet with the aliens. But then the question becomes: Will they be nice to us? How will they treat us? Will they disregard us or will they kill us with neglect? Will they use our resources? We have no idea and no experience of living with something a thousand or a million times more intelligent.”

“And that also amplifies the danger of sharing our planet with superintelligences that are in human minds. Augmented people with superintelligences. Ben Goertzel said Let’s get to an advanced AI quickly so that the first AI can be human augmented, but again we have no guarantee that will be nice either. In fact there is a certain probability

that it wouldn't be nice. If it came with our mammalian heritage for staying alive, with revenge, if it came with all that, then superintelligence would be very dangerous."

"If superintelligence would emerge from some complex system. How would you know that suddenly, well they'd probably use a lot of energy so suddenly you'd have a giant energy drain or it would disrupt some critical systems. The hard take-off from intelligence explosion, it doesn't happen gradually it happens very suddenly. I agree with that. I don't see why it happens gradually once we get to AGI, once we get to human level intelligence, then beyond that it would take a few seconds. It's hard to imagine something that is astronomically strong and powerful not being disruptive to our climate or planet or resources, the things that keep us alive. Just by neglect, not even by avarice or wanting to kill us for some reason or something just by not paying attention that there are other living things that are biological living creatures."

RICK CATTELL: "The old we would kind of be left behind if it would remain to be still just human. So, the new we would definitely be more bit than bio, yes and the humans left around would be left in the dust, I guess. I still think it's a far away in the future. I've been in Artificial Intelligence since 1978 when I did my PhD in AI, and look how little progress we made from then... We're making progress but in a very directed AI. I'm not sure we're anywhere near an AGI, where computers will be like people. Maybe in another thirty years."

"But in the sense of google becoming a superintelligence I'm not worried because I know how that software is structured and it's not structured to be super intelligent, it's just hardcoded to answer questions and it does that. It has a big search index, but it's not designed to be an independent intelligence. Google doesn't design other software. It's really very directed at answering questions to people. And they're not putting that much energy into AGI, they're putting more energy in google directing AI to the goal of answering a question. it would be a directed kind of AI."

ROBIN HANSON: "Fast growing worlds would produce a lot of fast changes that would be quite visible and noticeable almost anywhere and yes you would notice the difference. You wouldn't necessarily know the cause of that difference if it was a fast-

growing world.”

DAVID PEARCE: “There are mainly three scenarios. One is the intelligence explosion; Two is the Kurzweilian scenario of a complete fusion of machines with humans that may embrace mind uploading. Three, the posthuman superintelligence will be our descendants that embrace both genetic rewrites, recursively self-improving organically robots, but also and profited by *Neuralink* anything that narrow superintelligence zombie AI can be incorporated within you. So, you can have narrow superintelligence on a chip. Given my views on the nature of mind and binding I think there’s a sense in which classical digital zombies are invincibly ignorant. They don’t understand consciousness, the pleasure and pain axis, they don’t understand the nature of phenomenal binding, of course in all kinds of primitive domains they can systematically outperform humans. Nonetheless we can incorporate this kind of intelligence within ourselves, essentially it’s recursively self-improving organic robots that can encompass within ourselves everything that classical digital computers can do.”

RICHARD LOOSEMORE: “There’s a recursive bootstrapping where you have one single algorithm which is just satisfying some function that keeps operating on itself and overnight. And in that case, there’s no intervention by the engineers and they just pushed the bottom on the right algorithm, and it explodes and it becomes super intelligent by itself all the way up. And then there’s the version number two, in which people construct a superintelligence. They do all the work needed to get it up to near the human-like intelligence and then they duplicate it massively and get all those duplicates to help them, so in other words, they amplify their own capacity to develop the next version and then they all, the humans and these intelligences, which are not super yet, they are all working together to design the next generation...”

“It would be kind of an augmentation but not an augmentation of each and individual human brain but just an augmentation on a scale, you would just speed up the scale, speed up the work. In the paper that I wrote in the book (Singularity Hypothesis), with Ben Goertzel, the thing that we discussed most was the idea of number 1- amplifying by multiplying the quantity of computing power as I just described so you have many

duplicates. And also speed ups, just pure electronic speed ups where each individual intelligence in that set is able to have its clock speed increased enormously so that they are all perhaps only as intelligent as we are, but they're working at maybe ten, a hundred or a thousand times our clock speed. Those two factors combined are likely to lead to a situation where, as soon as we get relatively easy to build, say the size of a laptop pc, something that can be easily reproduced, as soon as we get an intelligence in such a situation, towards that task of building a new version of themselves. That will speed up but it won't be an overnight process.

Could it happen without us knowing it? Yes, in that first case. In the second case you would notice it pretty very much because there are a lot of people working on it and it's not going to happen overnight."

"I am deeply skeptical, about that first scenario even though a very large majority of the people who talk about the singularity just seem to automatically assume that the first scenario is not just possible, but likely."

Analysis #3:

This issue divides people between those that incline towards the so-called hard take-off of AI and those who think there is going to be a more gradual approach toward an AGI and a possible superintelligence stemming from it.

As Sandberg stated at the start of his answer, we could ask: what is the speed of take-off? We could measure it relatively to human reactions, but the hard take-off would be difficult to compare to human reaction times whereas in the more gradual takeoff, in which society and institutions change at the pace of its occurrence, as Hayworth duly underlines, we would witness a kind of gradual progression of which we would definitely be aware. Also, Koene acknowledges this and reminds that the laws of physics also affect that AI, so therefore any advance it should make would have to be quite visible. Hence, and as Hanson stressed, we would notice the difference, although in his reasoning this would not necessarily mean we would understand it, particularly in the case of a fast-growing world. But this is exactly what was implicit in this question, when it posed "would we be aware of it?", meaning would we understand it? Would we know it?

Hanson's answer seems a bit contradictory because according to him if we would not be able to understand its cause precisely because of the fast-growing world kind of rate we'd be subjected to, then how could we even tell the difference? Would we notice a change in the biosphere, in the ecology? Would we attribute it to every unexplainable Phenomena observed on Earth? Hanson's viewpoint can however be refuted. As Barrat claims, an emergence of an intelligence of that magnitude would be for us incomprehensible, so we would not be able to tell the difference, hence not even knowing it existed.

In his profound explanation, Sandberg contends that if superintelligence is to come about it is not going to be superfast in such a way as not enabling us to perceive it, because in order for it to be effective it will have to gather evidence, which takes time, although he acknowledges it might be faster than our reactions. And he goes on to state that we could in fact develop an outrageously intelligent piece of software so that we might not notice it. So, he admits the possibility. It is not that he believes in it, but nevertheless, he assumes that there could be this superfast kind of intelligence explosion we would not recognize. Barrat goes even further in his pro-hard take-off kind of scenario argument and claims that such a complex superintelligence would certainly require a lot of energy which could then disrupt critical systems and therefore would probably happen suddenly. So, he claims, an intelligence explosion would not happen gradually, but rather like an actual hard take-off. This is exactly the kind of argument that one would expect from the author of the best-seller titled *Artificial Intelligence and the end of the human Era: Our final invention*. Furthermore, Barrat thinks that as soon as AI would have really reached human level intelligence (so, an AGI), it would only take a few seconds for it to excel. To him, it is even clear that such an astronomically powerful intelligence would have to be disruptive inasmuch as deeply impacting our climate or planet resources.

In the words of Loosemore: "Could this happen without us knowing it?" Yes, it could, but as Loosemore sees it, only in the case of a recursive bootstrapping, in a hard take-off kind of explosion, in which one algorithm with a specific function would recursively operate on that function to then just suddenly emerge almost out of the blue. However, Loosemore is very skeptical on the likelihood of this happening. On the other hand, in the case of a gradual development we would certainly notice it because there would be a lot of people working on it and this would not happen overnight, although Loosemore still calls it an "intelligence explosion". As Loosemore argues, a variety of

different people would be involved in the construction of this superintelligence, joining efforts to get it up to near human-like intelligence and then they would duplicate it massively so that those duplicates would amplify their own capacity to develop toward a next generation, which he claims not to be super yet, but almost. He calls this a kind of augmentation but not one of each and individual human brain, just an augmentation on a scale. Loosemore recalls the time he was writing the paper “Why an Intelligence Explosion” for the Singularity Hypothesis book, co-authored with Ben Goertzel, and the things they discussed the most concerned an amplification by multiplying the computing power in order to get a large number of duplicates. This factor allied to the speed-ups, “pure electronic speed ups” as he calls them, where an individual intelligence of that set would be able to increase its clock speed, would eventually lead to an intelligence towards that task of building a new version of themselves. This would amount to a speed up, but it would never be an overnight process. On the contrary, it would take quite some time.

This approach described by Loosemore is similar to Sandberg’s view regarding an augmentation of some kind, leading to superintelligence. Sandberg mentions the Amazon Cloud as an example for this scenario. He claims that connecting various different intelligence sources with various different task-oriented skills is a useful kind of approach. Accordingly, we would achieve a sort of formal augmentation, which on the other hand might also mean that some people would augment themselves rapidly, which again is another worrying claim but he dismisses it quickly claiming that this would be rather a superintelligence as a service as opposed to a single kind of superintelligence, like an entity of some sort. Barrat on the other hand still finds ground for worrying scenarios in what concerns that augmentation or superintelligence. He stresses that human beings do not have an idea about living together with an intelligence far superior to ours, nor does humankind have the experience in dealing with such a situation. To him, this amplifies the danger of coexisting with augmented people with superintelligence.

As concluding remarks on this topic I would moderate general concerns regarding AI’s rapid and maybe somewhat ‘out of control’ progress with the views expressed by Cattell in the sense that we are making progress, but this progress is towards a very directed AI as he claims. He seems to be very much convinced that we are still far away from an AGI. And he grounds this view on the fact that AI’s research is more goals

oriented to answering questions, giving Google as an example for that. It is essentially a directed kind of AI. Finally, in what concerns the view expressed by Koch, according to which this AI is not an intelligence at all but only something that got better at copying, I believe it is very much influenced by his personal beliefs, or to be more precise, scepticism, since he positioned himself from the beginning as a fervorous sceptic of the singularity and issues related to it. For Koch there is no deep machine learning there is only deep machine training.

The sub-question of “Could this mean we’d be some kind of a cyber sapiens by this time?” was of course intentionally thought of because whether in a hard take-off kind of scenario or even in the gradual one, much of the perception would always depend on our own intelligence as observers. Of course, the interviewees essentially focused rather on the kind of intelligence explosion to determine if we would actually be able to be aware of it, which also makes perfect sense inasmuch as it is that intelligence that might turn out to be the disruptive one. But, if we were upgraded by that time, we might definitely be more adapted to such an event, thus perceive it and maybe still act accordingly in order to timely avoid an eventual bad singularity outcome.

As a general conclusion to this question, and given the interviewees’ opinions herein reviewed, I have to underline that guidelines and various approaches to the development of a so-called friendly AI have been discussed in the community for the last two decades. These concerns regarding a hard take-off, or even an intelligence explosion of the gradual kind, are of course covered in the SI chapter (Part IV). They also have some replicas on the following question concerning the rise of a singleton.

Table 12 – Selection of transcript excerpts regarding Question II – Q2.

Question II
Regarding Superintelligence and the Technological Singularity
<i>Q2: How possible would it be for a Singleton to rule the World?</i>
RANDAL KOENE: “Theoretically there is at least one scenario where you could play out where you end up with a singleton controlling the world. Whether that singleton remains a singleton or whether you can then start speaking of the various competing algorithms inside its own brain as almost being the population of itself as its own species, that becomes a matter of definition.”
“I actually don’t think it’s the most likely outcome because I think the world is very messy and there are so many actors in it.” —
ANDERS SANDBERG: “The singleton issue, which I think is a very important one because of the AGI scenario. That is relatively easy to turn into a scary story about how things can go badly wrong because the AGI is very powerful and it might have random values. The AI servers’ scenario on the other hand could also go very badly wrong because you can imagine evil people using superintelligence as a service to support whatever their evil plans are and there are lots of them. On the other hand, it might be that all superintelligences are owned by Jeff Bezos and his wife. So, it’s all in the amazon cloud. Nick Bostrom’s definition of a singleton is about agency. Who decides what’s going on in our civilization. So right now, we don’t have a singleton. We might have various kinds of corporate entities where we have group-minds like us, the voters and various organizations. And you could imagine a world run by the United Nations ++ (plus plus). Or google! Now, the interesting scenario here is: Suppose google could decide OK - we don’t want anybody to work on biotechnology that is too risky, we should shut down all biotechnological development. In this hypothetical world they could actually do that because they have enough power to do that, similarly the United Nations ++ or maybe the big AGI. In that case you would have a single point of

agency that could actually determine a large scale of things going on in the world and it would also be capable enough to say: Yeah, you tried to stop it but I can stop you from stopping it. The current mismanagement of covid-19 is a pretty good example. Inside China, its government is almost a kind of singleton and actually decides what happens. But we could imagine a world where we have dangerous technologies around and maybe we just need to have a system that if something dangerous shows up all the governments need to join forces in an effective way to do that, which of course sounds very utopian by current standards but might be something we can build the tools to build the tools for. At this point many people are going to be worried about how to be sure that this doesn't go badly wrong. A singleton is an existential risk too, if it has the wrong values."

DAVID CHALMERS: "Quite possible."

KENNETH HAYWORTH: "I do worry that the human race has never dealt with really extreme differences in intelligence. We dealt with extreme differences in Power I suppose but not in intelligence. I think that could happen, yes. The human race is built around this biological notion of a single self and there is no ability for us to really share different memories and to really incorporate different experiences into us and if we're talking about a world where MU has been perfected, then all aspects of the mind will be available for modification."

"So, I could absolutely imagine people sharing memories, changing their individuality theoretically. The idea behind being scared of this future and that we're going to lock ourselves away from a cure to all diseases including mental diseases; From a range of experiences that biological evolution simply hasn't given us access to; From space travel etc... Those advantages have to be sacked up against the scary aspects."

NICHOLAS AGAR: "I can envisage this. I don't know how long it will take before that kind of AI can write a better philosophy paper than me. I'm sure it's sooner than we think. Recently I wrote this paper called "Don't worry about superintelligence" and basically the line is yes, it is possible. So, Skynet is possible. So how worried am I about the singleton? That might say that I got rid of the other AI, like google because

that was turning into something, and humans made google, they're now my potential threat... I bet google could write me a great paper on Kant, maybe not now but soon, some google successor if it wants to. The hard things are easy, and the easy things are hard (Morovitz paradox). I mean the AI that controls our missiles and wants to form a plan to act on the world to achieve this end, that's quite difficult. So, I don't feel that scared of google one minute being a weather forecaster and the next it's launching nukes."

JAMES BARRAT: "Steve Omohundro talks about an ecology of AIs that grow roughly in steps together. There are two things. One is a scaffold of AI approach where you make an AI of limited intelligence to help you get to the next level of AI, and then you develop a better one where you create a scaffold to get to the next level, but never take away the scaffold and you always limit its power. That's one way to approach this. The other way, the ecology that a lot of AIs are maturing at the same time, and the theory is that they'll keep each other in check. None would grow so quickly or dangerously that it couldn't be stopped or slowed down by the others. AI will have basic drives of self-protection it won't want other possible hostile AIs to thrive in its environment."

"I like that theory but he's trying to impose some order on this unknowable thing. Again, it's an idea that is seeking for a religious paradise."

RICK CATTELL: "I think it is quite possible. Again, this is a danger of our "one small world". Over time, the boundaries of countries have become more and more transparent as we have evolved as a species. We are connected with faster and faster transportation, technology, and communication. All of that makes it easier and easier for one entity to rule the World."

ROBIN HANSON: "I could imagine that today! A singleton. And I could imagine a world government. We're not that far away from a world government being feasible or even natural."

DAVID PEARCE: “Though one can in some sense envisage some kind of global brain, I don’t think one can envisage global phenomenal binding. And in the absence of a unitary self, I don’t think one can envisage a singleton entity. How it is possible for an allegedly classically pack of neurons to somehow generate a unitary self is extremely controversial.”

—

RICHARD LOOSEMORE: “If it’s the overnight explosion recursive intelligence, the circumstances under which such a thing would arise would probably involve a very simple motivation system and, so if something arose through the night then almost inevitably it becomes a singleton. But you know, the trouble is, in a sense all bets are off because, all bets are off meaning it’s impossible to determine because if a superintelligence arises because of an overnight recursive explosion then the type of intelligence, the end product is going to be the result of a complex non-deterministic process, we don’t know what its motivations would be at the end of the rebuilding.”

“You can think of it in two things mostly: intelligence and motivation. [...] If the type of AI that is built is done is still an explosion but in the course of a few years and is done by people in the manner that I just described there is an immense amount of control and distributive decision making about how it ends up and just to summarize all of that is a book length, but to summarize it, all my feeling is that people would decide that actually they need a singleton or a cluster, a singleton cluster a redundant cluster and they need it really because once the general technology is available the one thing you don’t want is for people to go off into a dark corner and copy it and have the freedom to go and build some crazy version of something that was built very carefully by a very large group of people the first time around.”

—

Analysis #4:

Some people might find this question to be too far-fetched, perhaps a little ‘sci-fi-ish’. Nevertheless, some of the answers of these experts, especially those given by Sandberg and Loosemore show quite the contrary. Besides the fact that humankind’s destiny could end up in the wrong hands, the fact that these ‘wrong hands’, so to speak,

could be non-human, makes this an even scarier scenario. Furthermore, 9 out of 11 of the interviewees regard the singleton as a quite possible outcome. Consequently, I would argue that this proves that we need to invest in and develop ‘friendly AI’ in order to avoid the direst outcome.

Hayworth stresses the fact that we never dealt with extreme differences in intelligence in our history. But it is evident that with a rampant AGI, eventually out of control, this could very well happen. Cattell evokes the interconnectivity we are experiencing in today’s world with faster transportation, technology and communication to be a factor in making it easier for an entity to actually rule the world.

Agar envisages such a possibility. He imagines AI writing a better philosophy paper than he does quite sooner than we think. However, this does not lead him to worry too much concerning the dire outcome. As for Koene, he does not think of it as the most likely outcome, although he admits its possibility and discusses the definition of it in terms of various algorithms competing inside its brain, which in some sense connects with the conceptual view of Steve Omohundro here evoked by Barrat. They talk about an ecology of AIs that grow roughly in steps together keeping each other in check. Barrat says he likes the theory but underlines the fact that Omohundro’s trying to impose order on an unknowable thing.

Loosemore focuses on the motivation system. He thinks that in the case of a recursive overnight kind of intelligence explosion, as described in the responses to the previous question, it would certainly lead to a singleton. For such an intelligence to emerge in that kind of explosion it would surely involve a very simple motivation system. But in a case such as this, he also stresses that it would be impossible to determine because a superintelligence that would have arose out of such an overnight recursive process would be the result of a complex non-deterministic process so we would never know what its motivations would be like. AGI being friendly or not makes the singleton issue an important one, as Anders Sandberg duly emphasizes. The AGI might have random values, as he points out, and gives some examples of possible systems and companies such as Amazon and Google strengthening their power and gaining the decision-making leverage. He gives Nick Bostrom’s definition of a singleton as an example of a single decision-making agency that could really govern a lot of what is happening in the world. He concludes that given the wrong values, a singleton could be an existential risk, too.

Like the previous question, this one is also seen in the light of friendly AI kind of development discussed in the SI chapter (Part IV) of this thesis. The concern raised over this issue herein reflected by the interviewees' fear, mirrors that of the general scientific, technological and philosophical communities. Stephen Hawking (as quoted in this interview by some interviewees) also warned about a possible misguidance of AI toward a super powerful entity of some kind, in one of his latest interviews given to the BBC: "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last unless we learn how to avoid the risks" (Hawking, 2017).

On a final note I would recall Hayworth's claim of the previous question that fits perfectly: Our fears concerning all that could go wrong have to ultimately be overcome in order to evolve as a species.

Table 13 – Selection of transcript excerpts regarding Question III – Q1.

Question III
Regarding Transhumanism and Posthumanism
Q1: <i>Being human will be different in a Post-human Era. Could this mean we would definitely be more bit than bio by then?</i>
RANDAL KOENE: "I think ultimately that seems very likely. Some people claim that to create a proper processor, like a brain, you would need to build a neuromorphic chip that has basically an analog computing capability. That is a little bit silly, we have had these wars already, with compact discs versus analog music, you can always go to whatever degree of precision you want to, and it gets more reliable. And will we have more electronics or more software or more hardware that is developed and engineered versus evolved from biology? Yes, we will, because that can evolve faster and involve many more options. I don't know at what pace, but we'll be more bit than bio! I think it is likely that this is something that is coming up. Whether I would call it the next step in human evolution? I think I wouldn't say that because that confuses us. Some people will say, well theoretically anything we do is a natural step because we are a product of nature therefore everything we do is also natural. We still have selection pressures,

but these aren't the same as in the natural environment and it's not through the change of DNA because of natural selection pressures and our normal procreation. So, it's a step, in human's self-directed evolution."

ANDERS SANDBERG: "What we mean by natural. We are of course applying to it what I would say is a kind of industrialized or post-industrialized era concept of nature to it. We have been changing that. At the same time there is this important aspect of where is the route of your life, where is the foundation of existence. Many people definitely want to have a sense of good foundation for their own individual existence, "I'm a member of the biosphere and that gives me a sense of being part of something bigger than me". So, nature also becomes a point of allegiance. Many people use basically the great chain of being, that medieval concept of, God then humans then animals, then plants, then organs then minerals. There is kind of an organization, a structure of the world. And upsetting the structure is not just changing things. It's actually morally bad. And this is why I get a lot of resistance to ideas that change nature and of course especially transhumanism and posthumanism. Others would say "oh wait a minute, evolution is ongoing here, we are doing the natural next step here by modifying ourselves." By natural next step what they mean is they're trying to use natural here as a good word to depict that it's part of the natural order, which is usually implicit. Rational dignity of man is important for transhumanism. Humans got this fantastic ability to change themselves. And that is what gives us dignity. And that is what we usually do to get a concept of modern human dignity. I'm often briefing the case of human dignity to transhumanists and it's very fun because I've been arguing this with Francis Fukuyama that of course has a much more Aristotelian idea about human nature. No, modifying yourself means going outside the human, it can't be a good action for a human to become not human hence transhumanism is bad. There is a descriptive aspect of nature and evolution and where this allegiance aspect of tribe and a moral one."

"Evolution is actually an interesting process in that it invents new ways of evolving. Evolution has come up with various manners that change how evolution itself works. Mothers give some of their antibodies to their infants through breast milk. This is in some sense transmitting information between generations using a non-genetic factor, getting around the bottleneck of genetic evolution. And then of course nervous

systems with learning abilities get around another problem about updating neural connections and then a while ago human brains which are able to transmit information by talking. So, the interesting thing here is where do we say that natural evolution has stopped. I would say no, actually natural evolution is kind of discriptional how new forms evolve. It's just how the mediums get changed quite a bit. So, by this account I think getting to cyber sapiens in a post human era is totally natural from a descriptive standpoint. It's just what I would probably expect if I had a million intelligent creatures evolving on a different planet. The tribal allegiance. How different would you actually feel if you were a post-human. And that might be really complicated because there is one thing how you feel about your body and about your life. And then what story you tell yourself about that. So, if you think about brain emulation, in some sense it's the simplest way of making a posthuman you can imagine. You take a brain, scan it and you run it in a computer. And by assumption if this works you get the same mental states that you would have had with the original person. if I wanted to add various extensions and enhancements to my body in some cases I might want them to feel very different from how it normally feels, because that's the point. I want to deliberately extend myself because I want to feel transhuman. Sometimes I fear that we might split off into the windows and apple civilizations. We might have simple people that do upgrades but are not compatible for trivial and technical reasons. But then they become design reasons and that affects the way of your life and then of course both of them are going to have trouble in understanding those weird open-source Linux minds out there. So, what I'm hoping of course is that we get a lot of diversity. We can still hold together in a joint society because we have enough in common. There are always some people who want to isolate themselves and say we here are particularly particular for some reason, and that makes us better than others and that's going to cause all sorts of friction."

—

DAVID CHALMERS: "That's likely eventually."

—

KENNETH HAYWORTH: "I can definitely imagine that. The human race is going toward mind uploading. I think that's completely inevitable. The time frame is what is in question. I definitely would disagree with people that think that's going to happen relatively quickly. The human race will get better and better in scanning and

understanding brains and I see a moonshot level program for uploading the first human brain. That would maybe be able to be completed in about 50 years, but it might even be a hundred (100). So, within this century I think we will upload the first human person.”

“I’ve definitely seen projections of people that don’t take into account technology is on the horizon and how that technology will work synergistically to accelerate the whole package. Debating this with neuroscientists it becomes pretty clear that there is not a consensus on what level would be required for WBE and as for some of the levels that people suggest no matter how fast we’re progressing it would still take a lot of time. I am way less concerned about timeframes because my main stick if you will is brain preservation and we have the technology today to do brain preservation. If a civilization is way more advanced, it is more than likely to be technological. They have to undergo the transition of whatever their biology is to a new substrate that doesn’t have the same limitations. So, I really think that it has a logical progression.”

“It doesn’t take too much to understand that human beings are not going to willingly accept less than what they could get and that sticking with biology is always going to be less than what you could get by designing a new form.”

JAMES BARRAT: “Our intelligence has enabled us to survive. I think we can’t escape our biological roots. If we have an augmented brain, it’s going to be an augmented mammalian brain, and the core of that is a two million brain. And we can’t, we cannot transcend that! Even when all our needs are answered. We still won’t be satisfied. We always quest, we always seek.”

RICK CATTELL: “Doug Engelbart at SRI in the 1960’s talked about “intelligence augmentation” through computers that assisted a human in reasoning, collecting, communicating, and remembering information, and his vision has come to be, over the decades that followed. Medical advances have also allowed us to augment ourselves physically and mentally in various ways, and genetic manipulation may allow even more. A singularity of sorts will likely occur if mind-uploading becomes possible, and we indeed become “more bit than bio”, but that seems a long way off to me. “More

bit than bio” is a scary thought.”

ROBIN HANSON: “We are quite clearly enhanced humans relative to our ancestors. We have compromised in many ways to allow and embrace various kinds of enhancements and changes. There isn’t just a binary distinction. There are many gradations along the line from our ancestors millions of years ago to us today. And that will continue into the future. Whether we will embrace them as us is somewhat context dependent. When we look backwards, and we see the path that leads to us we see that we are willing to embrace all of that like us. When we look forward to things that we might become in the future we are less clearly willing to embrace that as us. Post-humans as you might call them will just call themselves humans. And they will look back and see them as a natural extension of their heritage just like you and I call even if our ancestors might not.”

“It’s wrong to state that the future is more dichotomous than the present or the past. It is a natural intuition people have based on something called construal level theory. Construal level theory is a natural way in which we switch between abstract distant things and up-close concrete things, and our habit when we go into an abstract thing is to think that it’s more dichotomous, that categories are more distinct, more internally consistent and more distinct across different categories. That’s just a more natural way to think when we think of abstract things and when we think about the future we tend to think very abstractly. There will be a continuum of enhancements in the future just like there is now.”

“Brain emulations would be more distinct from non-brain-emulations in humans and that difference will be the difference in the way people talked and reacted to.”

DAVID PEARCE: “Yes, essentially evolution via natural selection is proverbially blind and based on mutations that are random with respect to the direction of evolution. I think the very nature of the selection pressure changes when intelligence agents can actually choose the genetic makeup of their future offspring in anticipation of the effects of these changes. I think generally we have a very impoverished conception of the nature of intelligence.”

“Yes, it’s the next stage in evolution! Unfortunately, simply invoking these terms like transhumanism or posthuman doesn’t say much more than being trans Neanderthal or post Neanderthal...”

—

RICHARD LOOSEMORE: “We would definitely not simply remain just as bio. Bio is too fragile. Maybe we end up as some kind of hybrid of some kind of a positronic brain inside a biological body, a hardened positronic brain that can’t ever get damaged but nevertheless inside something that we’d prefer to keep as biology...”

“I think we’ll have the choice. I think a lot of people will immediately want to go and explore, you know to use their choice and to go and explore every imaginable variation, a lot of people would just go straight to silicon or whatever is the replacement for silicon.”

“But I think also people would be able to pour themselves back and forth between different substrates. Some people may choose never to leave the biology. There will be choice and there will be the possibility of making changes in that choice.”

—

Analysis #5:

Concerning this topic, it is quite interesting to find an almost unanimous answer to the simple but witty question: will we be more bit than bio? All eleven authors/ experts / interviewees feel that we will, in the years to come, become more bit than bio, including those who are *non-singularitarians*. The differences in their answers varied essentially over the timeframe, although even here most of them would agree that this process will probably develop later rather than sooner, which again is noteworthy. We can perhaps conclude that the interviewees are inclined to acknowledge that this will happen, but they are neither overwhelmed by enthusiasm for it nor blinded by it, which grants credence to their views on this subject.

Hanson refuted the cleavage I set when clearly differentiating post-humans from us, humans of today, given all the upcoming enhancements. The way he sees it, we are just a natural extension of our ancestors' heritage. Post-humans will call themselves

humans, just as we do with our ancestors and he grounds this reasoning in the fact that enhancements follow a kind of continuum in the future as they actually do today. It is also in our nature to feel part of something bigger than us and something we can relate to, which makes us feel somewhat particular, so nature becomes a point of allegiance, too, alleges Sandberg. He states that we have this ability to change ourselves, which in some sense defines us as humans, and which has also made him have some interesting arguments with Francis Fukuyama, whom he claims to have a more Aristotelian idea about human nature, which makes him (Fukuyama) state that modifying yourself means going outside the human, thus it can't be a good action for a human to become not human, and hence transhumanism is bad (Sandberg on Fukuyama). This in turn led Sandberg himself to conclude that there is a descriptive aspect of nature and evolution. As such, he goes on in his argument, evolution invents new ways of evolving and in this account it makes all sense to him to think of a cyber sapiens in a post-human era as being totally natural from a descriptive standpoint. Koene also addressed this aspect regarding the nature we are ourselves, as a human society, sort of stemming from nature itself. Theoretically anything we do is a natural step because we are a product of nature, therefore everything we do is also natural (Koene). And he continues claiming that we still have selection pressures but also that these have changed and are changing. He considers this to be a step, in human's self-directed evolution.

This is an argument that, with different variations, gathered some consensus among these authors. David Pearce also underlined that when intelligent agents are capable of altering the genetics of their future offspring then the selection pressures change. And he concludes this train of thought fully answering the question I asked, stating: Yes, it's the next stage in evolution (Pearce) [regarding a cyber sapiens]. For Kenneth Hayworth this question does not even deserve to be asked, given that this is just a fact. Humankind has already begun the race toward mind uploading, and he even says it is completely inevitable. To him the real question is actually just one of time. And on that particular issue he doubts it will happen as soon as many other authors believe it will be. In his time frame, he estimates the first human upload somewhere around the middle of this century (his timeframe, presented in a recent webinar can be found in this thesis' chapter of WBE-Part II). By the end of the century, he expects this technology to be commonplace, although he points out that this time-frame question is not really bothering him at all. For him, the timeframe is of minor importance, essentially also because we already do have today the technology and means to do brain preservation.

One of the arguments favoring mind uploading Hayworth finds, as already Koene had stressed in the preceding interview, is the fact that a civilization to be technologically competitive, adapted, has to transcend its biology (whatever the kind) into a substrate without the same limitations (of its biology).

Back to Anders Sandberg's rationale, regarding the transition to the post-human era, if someone would want to add extensions and enhancements to their body, then it would be precisely because they would want to feel different in some way, otherwise they would not want it deliberately. So, they would want to feel transhuman. This is the reason why Sandberg hopes we will get a lot of diversity in this post-human era, although still holding together in a joint society, as he says, because we still have enough in common for that to happen. On the other hand, and underlying Sandberg's vision of diversity we have his view on the insufficiencies of biology itself. According to Sandberg, to remain just as bio would not make sense because bio is too fragile and especially when we do have the technologies to enhance and endure ourselves. As such, and following this view, a lot of people would, also because of the novelty, immediately shift to silicon (or to silicon's future replacement), as Richard Loosemore elegantly described, but also, on the other hand he feels that there will still be people that will never want to leave the biology. Nevertheless, and still in his opinion, the majority of people would probably interchange between different substrates, as long as it would be possible to do it. Furthermore, and regarding the route to that sort of post-human cyber sapiens of some kind, Randal Koene also believes this to be a most probable outcome and he questions the reasoning of those who state that to create a proper processor, like a brain, we would need to build a neuromorphic-chip that would basically be able to compute in pretty much the same way our brain does. He refutes these arguments stating that we will have ever more software, ever more electronics and ever more hardware developed and engineered as opposed to evolving more from biology. And this simply because software and hardware can evolve faster and involve many more options than biology does. So, he just does not know at what pace but we will definitely be more bit than bio also in his view.

As a final remark it would make some sense just to remind James Barrat's words on this topic exalting the strength of our million year nature, and not underestimate its power: we can't escape our biological roots, if we have an augmented brain, it's going to be an augmented mammalian brain, and the core of that is a two million year brain

(Barrat). And still on that train of thought, Rick Cattell’s view also seems quite pertinent: a singularity of sorts will likely occur if mind-uploading becomes possible, and we indeed become “more bit than bio”, but that seems a long way off to me. “More bit than bio” is a scary thought. (Cattell).

Table 14 – Selection of transcript excerpts regarding Question III – Q2.

Question III
Regarding Transhumanism and Posthumanism
<i>Q2: Do you think politics in a few years will be divided between bioconservatives vs transhumanists?</i>
RANDAL KOENE: “I think it’s kind of there already. I think there is a possibility and it’s actually something we should prepare for or be careful of, because it’s already in the makings. The current politics mixes it up a bit and makes it hard to see, but if you look at it carefully you can see the little things of that separation, because ultimately when you look at what politics have been arguing about, it seems often about the pace of change. It seems that the more rapid pace of change causes a lot of concerns and groups of people anxious and it even threatens power structures that exist, and technological change tends to increase the rate of change or the pace of change in the world. A lot of the issues that play out in the political discourse are often ultimately based on exactly this sort of thing where this dividing shows up, and it’s already hidden in the politics today, and it is probably going to emerge even more as we progress...” — ANDERS SANDBERG: “Yes. Just look at many of the European green parties. They are left-wing bio-conservatives. I could imagine the future political debate between transhumanists and bio-conservatives, I think it’s not going to be that nice. After all,

why should environmentalists be on the political left wing? Indeed, historically they actually began on the very conservative right wing and it's kind of a weird coincidence that today you find most of the environmentalist movements relatively political left. And this has to do with compatibility between different ideas and also political alliances people make. You might see all sorts of these alliances. I think the real issue is of course shaping the view. Well what is the world like? What are people like? What do we do to give these views? And in this point transhumanism has an advantage in many ways because it's forward looking. Quite a lot of current political ideologies are very much based on past conflicts or past ideas. Basically, the traditional left-right argument has devolved into a question of how to move money around. The conservative vs liberal argument is more alive actually. Do we actually want to have social structures that are in some sense not defined by ourselves or do we get to define our structures in a flexible way. That also is linked to the nationalist globalist perspective but it's entirely possible to imagine transhumanists ending up on all sides because transhumanism as a movement of ideas has a weakness. It doesn't have a theory of a value internal to itself. So, the real question may actually be how much politics is going to be driven by this and how much is going to happen also with other non-politic players like cultural technologists. And different transhumanists give relatively different routes. Most transhumanists that you would find today of course have a fairly liberal view: oh, morphological freedom is very important, you need to be able to control your own body, others should not be able to control how you upgrade yourself etc... so the government has to find what your body's like, that sounds even worse. Of course, bio-conservatives, their solution, if we just avoid developing these technologies, then it would be fine, because then the question doesn't come up, which of course misses the point that these technologies already exist. The bio-conservatives can't win that one."

"Socialism actually killed much more poor people than capitalism. The real story here is how new technology spreads unevenly. But when the price of a technology comes down, gadgets like my smartphone are super advanced. It's about technology diffusion, the price came down exponentially and became democratized by diffusion. The price of drugs can come down exponentially once patents expire. Services on the other hand tend to be really expensive. Why? Well if you need to go to an enhancement spot to get upgraded you need to pay for the doctors and the nurses' salary and of course your own time-out from job. That also is costly. So, enhancements

are kind of service based. They are more problematic from an equality standpoint than the gadgets. I think we have an intuition here that maybe upgrading your body will be more like a service than a gadget. That is an interesting question which technically one can properly analyse separately for different forms of enhancement. When it comes to where power resides in society a lot of it has to do with where does the benefit go. On the other hand, many governments would dearly want to enhance most of their citizens. Just imagine if you could boost the immune systems of your citizens. They have less sick leaves, they are more productive and don't get hit by pandemics and they're probably going to thank you for having a great health care system. Indeed, many forms of enhancements might be just the thing governments would love to subsidize. If it turns out that this is good for people and the economy then they would want to do it. Which is of course why governments are subsidizing education rather than saying oh no let's just have billionaires run the universities. Instead we actually use tax pay money and make schools mandatory, which I think is really an interesting example where we made one kind of enhancement into a mandatory thing. You can't be part of our society unless you can read and write. So, I think we are going to see these debates showing up again because it might very well be that in fifty years that governments are going to be very concerned if parents want to have kids that don't have genetic upgrades."

—

DAVID CHALMERS: "There will certainly be such a division, but I have no idea how important it will be in politics (probably important at some point, at least for a while)."

—

KENNETH HAYWORTH: "I would like to think we're progressing very quickly but at the same time the human lifespan is very finite."

"Politics at some point in the future will be those who want to go faster into Brain Computer Interfaces and mind uploading and Intelligence augmentation, artificial intelligence in terms of incorporating data to our daily lives (general intelligence)."

"There will be people that want to do that and people that have seen all the movies that tell you that's a bad thing to do and try to slow that down."

“Right now if a politician said we need to really worry about AGI may be taking over the world in a few decades everybody is going to laugh because we don’t have artificial general intelligence and that’s not really a debate for right now but will be probably sooner than most people would probably think - but not right now.”

NICHOLAS AGAR: “I do think that when it comes to the bio-conservative rejection of human enhancement... But that ship already sailed. It doesn’t even make sense to sort of say oh I’m going to reject all the enhancements, because many of them are just part of our ordinary existence.”

“You know the old comparison of genetic enhancements once we achieve this how different is that from educational enhancement. Enhancement is part of our culture. The interesting thing is that it makes it difficult to work out which ones we think we want and which ones we don’t. I would like that to be a collective decision.”

TOM KOCH: “For a lot of these things it’s not about an enhancement but a prosthetic. What we already have is a dystopian divide in every level. I see no need to put this in the future that I don’t believe in when I see this happening in the present. I think that in terms of ethics we need to recognize that these are in many ways inconsequential in the immediacy of the day and that the problems that you posit to the future are concrete realities of the present when we understand the limits.”

“I think it is important (*in reference to me interviewing MU detractors*) and I congratulate you to have not only those super salesmen of the future. But also, the critique of them.”

JAMES BARRAT: “Why dystopia? Let’s look around ourselves and we look at the corporations that are in charge of this intelligence explosion. Facebook is just a huge manipulator of privacy and stealer of data, dishonesty to the core. Google totally corrupted by its vast bottom as well. Apple uses slave labor in China, among other things. Corporations are almost by definition corrupted at core. Take these corporations, take their ethics into the future. The future is governed by the cyberworld, and that’s a realistic argument extrapolation of what the future looks like. Regarding

politics, I think by the time of transhumanism there will be no politics anymore. I think any notion and representation will be gone. My challenge to transhumanists is what are you basing this utopian vision on? What period of Man's history has approached anything like this? Or sustained itself?"

RICK CATTELL: "Yes. We are already politically divided on that."

ROBIN HANSON: "The odds that the difference between being enhanced and not-enhanced would be the main political difference – just because it's one of many great differences – is not obvious. Why should it be that one that becomes dominant? Therefore, why would augmented versus not augmented, or computer versus not computer or more tech versus less tech, why would that be the basis of politics in the future? So similarly, the idea that there will be the enhanced versus the non-enhanced, well that's an elite versus a non-elite sort of, so it's just unlikely especially for a main political axis."

DAVID PEARCE: "In practice, if some truly revolutionary breakthrough in anti-aging in medicine occurs then essentially, I think that all the accumulated religious traditional bio-conservative objections to the idea of immortality and radical life extension would simply dissolve. But in another sense transhumanism cuts across the traditional left-right divide. The very nature of any information-based technology is going to be true with virtually anything that can be digital which means most things."

RICHARD LOOSEMORE: "In the middle of the sixties [...] there was one big divide between new style rock & roll and old-style rock & roll. The one big divide, which is a sort of like the left-wing right wing. So, the people on the right liked the music of the fifties and the people on the left liked the new loud brash rock roll that would break all the rules. Now, today, what do we have? There are no pop charts anymore. There are, but they're heavily managed. Everybody listens to something totally different and the music scene is fractured. [...] You could name fifty (50) genres now probably that people really exclusively are interested today. [...] I think that is representative of what will happen in the future as far as politics goes. So, the division between the bio-

conservatives and not will actually be a big spectrum. You will be able to name 50 different sets of opinions about that. But the whole question about how politics itself, you know decision making will play out, is going to get kind of fuzzy because I suspect that society will break into domains that want to live a certain way and they will basically appeal to the AI, people will probably like the fact that they can delegate a lot of stuff to the AIs. [...] People will fracture into certain groups that want to live certain ways but one thing they'll find out is that the tradeoff for having that ability to live the kind of life they want with the kind of people they want will come with a tradeoff that they won't be able to go to and impose it on anybody else. And I think they will live with that trade off and as a result politics will become very different because there won't be bio-conservatives arguing... [...] Some people will stay bio-conservative for a long time and then some of them will allow themselves to expire and allow themselves to die, because they don't care or they don't want to. I think many of them will have a last minute conversion and say "actually I'm a bio-conservative but now I got cancer (or whatever) and I actually would like to have my body changed and have the genetics changed in such a way that I don't have this, and they will appeal to the AIs and all of the sudden they will become bio-conservative but not quite."

Analysis #6:

Also, on this topic one can acknowledge a general agreement on the idea that politics could become divided between transhumanists and bio-conservatives generally speaking, with the exception of only one author (R. Hanson). Whether because of the power of technologies to be and their consequential rate of changes in the pace of things, whether because of the increasing trend of enhancements themselves. Furthermore, those that don't envisage such a scenario as being likely, mostly do so because of the eventuality of transhumanist technologies in the future already having edged or broke through in such a manner as to dismiss any eventual opposition to them.

For a politician to debate this issue nowadays and to worry about AGI's dominance of tomorrow would be almost ridiculous because there still isn't any AGI

visible today. But as Hayworth claims it will probably be sooner than one might think. David Chalmers underlines that the division will take place although he does not know what importance it will entail to politics. He thinks that it will actually be important for some time. Technological change tends to increase the pace of change in the world, and in turn this seems to cause a lot of concerns and make people anxious, even threatening power structures. For Randal Koene, it therefore is pretty natural to argue about this possibility in politics, since politics also has a lot to do about the pace of change. He thinks it is already in the makings so we should be careful of it or be prepared for it. Furthermore, and just as Nick Agar refers, enhancement is part of our culture, and to reject all the enhancements just doesn't make sense. Many of them are part of our ordinary existence. So what does make sense is to discuss and debate it in order to decide about its distribution, its diffusion, and make that a collective decision, as Agar stresses.

Richard Loosemore makes an analogy with the musical pop scene dualistically divided in the sixties to show that today there is a much diversified spectrum of musical genres where one could find itself choosing one of fifty different options. He thinks it is representative of what the political mainframe in the future could look like, thus there will actually be a big spectrum inside this division between bio-conservatism and transhumanism. Loosemore thinks we will be able to name 50 different sets of opinions about that issue. He suspects that society will break into domains that want to live a certain way and that they will appeal to AI, even delegating AI with a lot of decisions they do not want to make. Ultimately, he states that politics will become very different and although many bio-conservatives may stay this way for a long time, they will eventually end up breaking down just because they too will be appealing to this AI and will be needing and using the technologies themselves. This is a view somewhat shared by David Pearce in the sense that he also thinks that it would only take a revolutionary breakthrough in anti-aging or something like that where all the accumulated religious traditional bio-conservative objections to the idea of immortality and radical life extension would simply dissolve. (Pearce).

James Barrat goes even further in his view, claiming that there will actually be no politics anymore by the time transhumanism rises. He rejects the dystopian kind of scenario by exposing all the major Intelligence Corporations, which are in fact in charge of the intelligence explosion, as being corrupt and absolutely deprived of any kinds of

ethics. He fears those will be the intelligence agents of the future, since the future will most certainly be governed by the cyberworld.

Anders Sandberg predicts a future political debate between transhumanists and bio-conservatives. He also foresees all sorts of alliances which have more to do with the compatibility between different ideas than with the parties or their positions in the political sphere itself. He thinks that it is all about shaping the view, where he thinks transhumanism has an advantage since it is forward looking. Nevertheless, he imagines transhumanists ending up on all sides because they have a weakness as a movement of ideas. He claims that transhumanism does not have real theory of a value internal to itself. This is where other non-political players, such as cultural technologists, get in the picture. Furthermore, he also dismisses some of the bio-conservatives' power in that these enhancement technologies already do exist, a view that many of the other interviewees share. David Pearce also mentions the fact that in some sense transhumanism cuts across the traditional left-right divide. He states that this is due to the fact that information-based technologies are becoming essential transversal in society since most things are or tend to be digital

So, it's all about technology diffusion, the price comes down exponentially and things become democratized by diffusion. But enhancements aren't really gadgets. Enhancements are kind of service based, as Sandberg says, and so upgrading our bodies will be more like a service than a gadget kind of thing and a lot has also to do with where the benefit goes, in terms of where the power resides in society. So Sandberg argues that governments could gain a lot by enhancing their citizens just as they do today with education, which basically is a successful kind of enhancement that governments have put through in most countries, so they might even be subsidized, if proven to be beneficial to the power in the state of course. The transhumanist perspective to benefit from new technologies (i.e., enhancing human capabilities, etc.) remains highly controversial, as far as to be called the modern world's most dangerous idea by Francis Fukuyama. (Filip Bardziński, 2015). However, today it represents the vision in which our biology serves as a platform for launching a range of technologies, thus extending our natural capacities. (Fuller & Lipinska, 2014).

We are approaching an era where human enhancement is starting to be addressed as the norm rather than the exception, which in turn opens a whole new debate over the meaning and the significance of post-humanism also with a particular

revival of eugenics. This philosophical-anthropological approach to the delineation of human enhancement paves the way for discussing the norms concerning who counts as human (i.e. humanness) and also enables a framework for apprehending the intimate relations between humans and enhancement technologies. Also, the ethical frameworks we rely on such as utilitarianism, rights and autonomy would be seriously compromised. Barrat repeatedly claimed in his interview that transhumanism has religious motives behind it. But whether the belief is motivated by a religious impulse or not, and although it is naturalistic but opposed to an ethics of natural law, it does follow an ideology of pursuing human improvement. (Hopkins, 2005). Or at least it claims to be a class of philosophies that seek to guide us towards a posthuman condition. (More, 1990, More, 2011). The debate between bio-conservatives and transhumanists over human enhancement, and the directions to be taken in its respect, is certainly expected to grow in politics in the years to come. Also, as bionic prostheses, implants, neuromodulation devices, nano-pills and chips, and engineered tissues gain momentum and become common, it will in fact be ever harder for bio-conservatives to resist in time, as also some of the interviewees (Sandberg, Loosemore, Koene and Pearce) underlined in this particular question. However, Don Ihde, american philosopher of sciences and technologies (not interviewed for this thesis), alerts that far too often transhumanists fall into what he calls the Idol of the Cyborg — where technology would become completely ingrained into us and be completely transparent, but points out that no technology will ever be completely internalized, and that transhumanists are too concerned with merging machines with the human body. (Ihde, 2011). If we take the interviewees' position held in the previous question about us humans becoming more bit than bio in the near future for granted, then not one, but many technologies might actually be completely internalized in the post-human (check the plot-reprise at the end of this chapter).

In conclusion, Sandberg's view kind of reflects an overall thinking conveyed by most of the authors in respect to the possible bio-conservatives vs transhumanists clash in a near future ingrained by these debates because and quoting: "It might very well be that in fifty years governments are going to be very concerned if parents want to have kids that don't have genetic upgrades." (Sandberg).

Table 15 – Selection of transcript excerpts regarding Question IV – Q1.

Question IV
Regarding Superintelligence and mind-uploading
Q1: <i>Would it make sense to iterate human intelligence via mind-uploading to augment it / improve it, and in this manner cope with AI's progress?</i>
RANDAL KOENE: “There are some problems as well as some brain regions that you could improve well through such a simulated evolutionary approach rather than through direct engineering, because maybe you don’t know the direction you are supposed to engineer, or you don’t have sufficient hints to tell you where you’re supposed to go with it or you don’t know how you would make the changes that don’t make sense because you don’t understand the system well enough. Those are all reasons for applying such an evolutionary approach such as that constant iteration where you would get some kind of random distribution of possibilities. If you have a better understanding of the system and how it works or if you have a better understanding of the problem and how to optimize towards a solution then often a more engineered approach is the faster one, even if subsets of that kind of engineering involve this kind of, what you might call deep-learning experimentation or a simulated annealing or whichever approach we would want use to apply a little bit of random to find the best path. From an engineering perspective you would have to choose carefully when a certain method is going to be your most effective approach to an improvement in the output performance. Apply that only to unconscious pieces of a brain and never to an entire whole brain emulation, because if you believe that you are real when you are a copy, in other words, when you do not philosophically think that that is a nonconscious, non-thinking non-identity then subjecting all of these uploads to various scenarios is the same thing as subjecting people to them. That would be extremely unethical. So, there’s the ethical side and the other side is the engineering consideration. But yes, sometimes an evolutionary approach is good –

sometimes – not optimal.”

ANDERS SANDBERG: “So, I think this makes sense, yes. I quite often think that it would be quite useful and be able to try modifications if we exist as a brain emulation. You might change some receptor potentials or other things, and then after a while you test whether this was good and if it wasn’t good you can always undo the change and revert to the original. Of course, you can do it like you described and have these kinds of parallel evolution. It might be that you’d try out a hundred versions of yourself and see which variation helps the most and then you’d keep that one! Now, a lot of it depends on your view about what you do about spurs, these small copies that might not have a long future. So, for example, Robin Hanson [The Age of EM] thinks that most people would be kind of happy to have a lot of spurs and delete the spurs. I don’t think that’s true for most people. I think in theory I’m ok with it but I’m not certain I would actually be ok with it if I was in that situation, but it’s worth considering. There is this issue of how much are you allowing yourself to use versions of yourself as a resource. There are ethical issues around it, but I do think it’s potentially quite powerful in making new modifications especially if you have opaque systems. Brain emulations are essentially black boxes. It’s a model of the brain and it has all the messiness and we’re probably going to learn a lot of neuroscience once we have brain emulation because we can do a lot of experiments including neuroscientists that can do experiments on themselves and undo their effect, and then report back on what happened. So, this is powerful for black box systems. If you have something AGI based in some sort of understandable code and you can read the code and prove things, you get a very different form of improvement path. The question is which one is going to win, and I have a feeling that it might be easier to improve AGI than human brains because we have a backwards compatibility issue which is pretty tough. I think combining human brains with AGI will be very powerful but now you need to think they are very, very different systems and that’s going to be a tricky task. So, I have a feeling that in the long term, I have my doubts that humans are going to compete even if you try to upgrade them. I might be wrong it might also be like Robin Hanson’s suggestion and human-like minds might be part of AGI systems.”

DAVID CHALMERS: “Yes. Maybe, but there is still a danger of singleton augmentation. Would Universal Darwinism not select AGI as the more adapted intelligence instead of ours? Yes, though I’m not sure that putting this in terms of Darwinism is helpful.”

KENNETH HAYWORTH: “I think that people will definitely do that, yes. But I think the key question here is what level of consciousness is required for general intelligence. Our conception of a mind upload is essentially a human being but emulated in a computer. So that package of the human mind is more useful than any artificial intelligence today is way lower than human beings because we have this general intelligence (common sense and stuff like that etc...). So, when you’re saying to iterate that, that could be a very, very powerful augmentation because now you have a thousand “yous”, momentarily, working at the same time for the same problem potentially from different aspects. This of course assuming that whole brain emulation can be made relatively inexpensively, which is a big assumption, but at some point, this will be the case.”

“How much of our common sense and general intelligence is tied to our consciousness? Can we have the type of really subtle common-sense intelligence without bringing in things like suffering and the sense of self and things like that. And if I had to guess, I suspect that those are decoupled, so iterating under that scenario where you would decouple your general intelligence, you would never want to iterate yourself because it’s inefficient.”

“There’s a type of people that think that WBE is going to require a simulation at the ion-channel and receptor, molecular level and somehow the human brain is using the conformations of individual proteins that are absolutely fundamental to the algorithms which of course are absolutely fundamental from an implementation level, nobody would doubt that. But (...) there are people that think that there really is a crucial computational level, there are people that would think that a human brain boiling down its algorithms, its algorithmic complexity might come up with something that could be a regular sized supercomputer today versus people that would think that it would take billions of supercomputers mapping out the conformations of individual proteins.”

“Algorithmic complexity is not very large to begin with and therefore I think iteration is something that is going to be doable sooner rather than later. Sooner might be in a hundred years...”

“We could emulate a brain really well if we’d understand how it works. Our computers are capable of the algorithmic complexity. We just can’t program them. I think that we will understand the brain before we get to the technology to emulate it. I think we will reach a completely satisfying algorithmic understanding of the brain. We wouldn’t have to. But I think we will.”

—

NICHOLAS AGAR: “It’s a great thought experiment! You described enough about it to make clear that it does not violate any law of physics. It’s potentially doable.”

“A thought experiment like that, is one of the defining features of the digital, copiable, designed to be precisely modified in a way that in principle we can modify ourselves. And we can presumably make the changes you described, in a way we try to work out the best digital version of *you*.”

“In a way there’s a lot of versions of me and let’s put them in a situation like climate change so that they can interact with each other and interact with intelligences that are copies of you and that are capable of engaging this sort of decision-making that will promote welfare and are also not going to cause so much carbon into the air. I mean imagine what that could have a potential approach to these crises.”

—

JAMES BARRAT: “That sounds feasible and there’s some speculation. And yes, well you basically described evolutionary algorithms. I think all that sounds promising. It’s a tower of suppositions that can be successful. It’s an interesting thought experiment. It’s a good thought experiment. If you develop some of these cognitive architectures in the virtual world you avoid some of the safety issues in the real world.”

“Ben Goertzel, for a long time, and a lot of people, have been talking about creating a virtual world with very, very high resolution where you have (to navigate in that world you would have to have the same ontology) common sense that you have in the real world, and then raise them through different iterations, better and better architectures that can navigate that virtual world. And here’s the question: How do you introduce

ethical education to that creature? But that [in reference to the TE] seems more plausible than raising an AI in isolation and trying to program it with ethics. It seems like if you could raise it in an environment where ethics, where kindness and other virtues were rewarded then you would have this hard-wired nice thing instead of letting all the commercial drivers evolve it.”

“There’s a lot of promise in the idea of developing things in the virtual world. Which is really what you’re talking about. About different brains exploring up in a virtual space and then experimenting and developing them in that virtual space.”

RICK CATTELL: “Argument in favor: We have the benefit of starting with the best of billions of variations on human intelligence, and an intelligence evolved from humans might be better at communicating with humans and helping to satisfy human needs. Argument against: We can better predict the goals and control the behavior of a designed intelligence “architecture,” and we can produce intelligent behavior sooner, indeed we can already demonstrate levels of AI in limited domains. Yes, you could do that.”

“That would be sort of fast evolution wouldn’t it? Iterating on yourself and in a virtual world so you wouldn’t need to worry about exhausting the resources of the planet. Well it’s not technologically feasible for the time being and for quite some time, but if you could do that in a virtual world you could create fast evolution.”

ROBIN HANSON: “The Em economy [in reference to Hanson’s book *The Age of Em*] would be very competitive. You would make different versions of yourself to compete in different parts of the “Em economy”. Some versions would succeed more, and others would succeed less. And then you would end up making more copies or versions of the ones that succeeded more that would be just a natural result of the economic competition. So you’ll get old and no longer be able to be as productive so you’ll keep having these younger versions of yourself and so that they can move toward a subjective age of peak productivity after which they would wear out, make modifications in the emulation as best you can to understand what works well in the context of the jobs that you were going to give them. And so be part of the process I call those tweaks in the book, that there will be a search in the space of tweaks. And

my prediction is initially sort of surface tweaks, tweaks you don't understand very well that you try, and then you will have wins. But those wins will run out quickly, and they will sort of be stuck in the tweaks that you can find and you're probably not going much further until you really understand the brain better. And so, there will be this long period where we just don't understand the emulation brains. So, the simplest way to modify will be to look at it and say most jobs people need to do, don't need the high resolution of the sight and hearing that the human brain has. So, you know what? We're just going to take out a lot of that and end up with an Em that is still capable of doing its job and a lot cheaper because it's not going to have all the brain devoted to it. So the idea is you have the spaghetti but you will try simple variations and some of them will be wins and that will be the early Em where you focus on these simple tweaks that you don't require to understand the brain in detail it just requires that you play with things and see how they work."

"Honestly that's how most innovation has always gone in history. Most innovation just hasn't come from some deep understanding of things, most innovation comes from just trying variations and that's what we'll do with brain emulations."

—

DAVID PEARCE: "The background assumptions aren't unreasonable given Turing machine functionalism but they're precisely what I would challenge in the sense that if you'd want to have exclusively biological superintelligence I would start by cloning Von Neumann and essentially hothouse the clones, tweak them, put them all in the program and this recursive cycle of self-improvement."

"Now, plenty of people would say "Yeah biology is slow and it's primitive, it's carbon chauvinism, we need to escape our primitive substrate". But they have to give an account on how to solve the binding problem. The hard problem of consciousness, they have to explain the causal efficacy of consciousness (...)."

—

RICHARD LOOSEMORE: "If you are interested in testing lots of different kinds of intelligence mechanisms, then whether they derive first of all from an emulation download of your brain or from an AI design that you have, that never came from a human being but just involves that other mechanism; If you're involved in just wanting to do lots of Darwinian testing on different variations on the design you can either build

a complete system, so in the case of the brain emulation of you, a complete version of you [...] and then you do that a million times, a million simulations all in their own separate virtual laboratory each of them tries to work out the outbreak scenario and each does it in a different way and you select the best solutions. [...] we'll call that the full system test, but you could also do it in a different way. If I understand which part of the brain is doing the actual thinking, the core thinking part of it, we could just take that part out and it has no longer the rest, so there is no longer a consciousness. So, this is a unit test. This is a subsystem test and you don't have the danger of building a consciousness. But you only test the thinking parts of it in a very tight environment where you don't give it a lab to look at, you put in patterns and it relaxes and comes out with these solutions in patterns but it doesn't think about itself, it has no self to think about because that has been all taken away. [...]"

"I think they'll have to get the whole brain emulation thing to work in the first place, if they understand enough they would do subsystem tests on the mechanisms inside and to the extent that they use your mechanisms, your biologically developed intelligence mechanism, it will look like you, it will think in the same way as you but it won't be you, they'll make sure that it has not the life memories that you have, personal life memories, background knowledge perhaps, but not episodic life memories (in psychological terms). So, I think that is feasible and you could try to do it. [...] But you know, that kind of thing will also be a key part of a superintelligence explosion of the type 2, which takes a few years. That would be an absolutely core part of it. You've described exactly what will be going on most of the time during those couple of years of development. Much repetitions, many variations, experimentations, will be on the core mechanisms of intelligence."

Analysis #7:

In spite of the obvious similarities between the thought experiment developed in this thesis on iterating human intelligence via mind uploading and the "Em" idea described by Hanson in his best-seller *The Age of Em*, there are significant differences

between them that justify my own thought experiment. All interviewees seem to worry about the philosophical and ethical issues linked to this thought experiment's possible consequences. Regarding personal identity and consciousness, in a first approach in section I, and when confronted with the question "would another me be me?", the answer of the majority of the authors clearly indicated some doubt concerning this possibility in the sense that there wouldn't be another me exactly as me, since the elapsing of time provides different experiences for all the copies. Now, once confronted with the idea that other copies of me could be iterated in order to obtain my best me, ethical grounds were evoked for the possibility of actually dealing with me, myself and my consciousness. This seems to me a bit like first arguing that in principle MU would not actually be able to successfully produce fully conscious uploads when analyzed as an eventual means to achieve it, whereas when we assume it to have actually already happened, then suddenly we have to be very careful because conscious copies might have been produced after all. It is a bit like the tele-transport analogies made in order to evaluate the uploading scenarios. As Hanson exemplified when talking about the construal theory. Once presented the scenario to the students as before entering the teleport (regarding the future), the majority would dismiss the possibility of success inasmuch as claiming the upload (or teleported) person not to be themselves (or possible). But when regarding in retrospective, so assuming teletransportation (or upload) did actually take place, ironically everybody in that student community felt like they were themselves, inasmuch as stating that a successful upload (or teleport) must have happened, since I'm still feeling myself.

Randal Koene alerts to the fact that there are ethical and also engineering reasons to consider such an approach. If we don't understand the system (the way the brain works) well enough, then this kind of evolutionary approach would make more sense than through direct engineering. But otherwise it might be better to avoid such an evolutionary simulation or deep-learning experimentation, as he calls it, simply because we might be uploading a conscious system and submitting it to experimentation. Rick Cattell comments that by iterating myself in a virtual world I wouldn't need to worry about exhausting the resources of the planet. This isn't entirely clear to me, as uploading, with all the investment it involves, would probably spend a great deal of resources. However, his time-frame in achieving WBE and MU is along with Kenneth Hayworth's, a rather pessimistic one (on a stakeholder's kind of view) claiming: It's not technologically feasible for the time being and for quite some time, but if you could do that in a virtual world you

could create fast evolution. (Cattell). Ken Hayworth thinks the level of consciousness required for general intelligence is a key factor for such an approach. To iterate our uploads could be a very powerful augmentation as Hayworth underlines: because now you have a thousand “yous” working at the same time for the same problem potentially from different aspects. (Hayworth). But the question remains. How much of our common sense and general intelligence is tied to our consciousness? So, just as Koene had hinted at in his approach, dealing with conscious parts of the brain, so to speak, is pretty much like dealing with persons. Hayworth, nevertheless, suspects these two elements (general intelligence and consciousness) to be decoupled. But as he admits this is just a guess, based on his experience and knowledge but nothing more than that. Nevertheless, Hayworth doesn’t think that a WBE would require a simulation at the ion-channel and receptor, and molecular level, as opposed to other authors who think this way. He argues that these other authors who think this way believe that somehow the human brain is using the conformations of individual proteins that are absolutely fundamental to the algorithms, crucial on a computational level aside from being fundamental from an implementation level (and he does acknowledge that they are fundamental at this last level). But he argues that they are not fundamental at the computational level, just as other interviewees, such as Richard Loosemore, who also emphasized this point in our interview. Hayworth adds that algorithmic complexity is not very acute, i.e., it does not pose such a big problem, and therefore an iteration of this kind could be doable sooner rather than later (although sooner in Hayworth’s timeframe could mean a hundred years, as he says): “I think we will reach a completely satisfying algorithmic understanding of the brain. We wouldn’t have to. But I think we will.” (Hayworth).

To James Barrat the thought experiment I put forward does sound feasible even though it involves some speculation and entails a lot of claims. But they can prove to be right, and he explores the possibility. In doing so, he compares the thought experiment to evolutionary algorithms in the sense that one could develop these cognitive architectures, as he calls them, in the virtual world and thus avoiding safety issues of the real world. The suppositions behind the thought experiment also pose a problem to David Pearce, not given their speculative level but because those are precisely the background assumptions he would challenge, although he doesn’t find them unreasonable given Turing machine functionalism as he says. He would challenge them because of the binding problem and the hard problem of consciousness that he already had expressed

to be his major concerns. To have exclusively biological superintelligence one has to solve the causal efficacy of consciousness.

Richard Loosemore describes the scenario put forward in this thought experiment as an experiment that involves testing lots of different kinds of intelligence mechanisms that were emulated from my complete self to do a simulation of millions of Darwinian variations of it. To this process he calls the full system test. He suggests though, that this could be done in a different way, which he would call the subsystem test in the sense that it wouldn't entail the parts of the brain responsible for consciousness, so to speak. He would take out the part that does the actual thinking, the core part of it, assuming he would by then know which part that is (if it were to be a part, which I suspect people like David Pearce would disagree insofar as the binding of a unitary self would clearly be affected), in order to detach the danger of building a consciousness. So, these parts would not actually be able to think of themselves since they wouldn't have a self to think about. One of the pre-conditions for this subsystem test to work would of course be the fact that WBE would be totally understood without any kind of glitch otherwise you could always have those near-misses Nicholas Agar was talking about when answering the previous question of the interview. Loosemore goes on to conclude that this kind of approach is essential in the superintelligence explosion kind of scenario he described earlier, the second one, the one that doesn't occur overnight, and that would take some time to develop. He finds that the thought experiment herein exposed fits perfectly into that kind of algorithmic development toward the superintelligence, a development filled with variations, and in which repetitions would define the core mechanisms of intelligence.

Robin Hanson mentions the similarity of the thought experiment I propose with his Ems, in reference to his book *The Age of Em*. He then approaches the thought experiment scenario by describing the Em economy and detailing some aspects of it. Just as in any kind of evolutionary approach, his successful Em tweaks are specifically designed emulations to operate a specific function in direct relation to brain functions and regions. These wouldn't be the result of a WBE and so they would clearly avoid the problems outlined by Loosemore, Hayworth and Koene already mentioned above, and therefore bypassing the consciousness problem. So the best tweaks would be competitively selected, as would the uploads in the referred thought experiment, and then of course improved in the sense that, and still referring to this Em Economy, most jobs people do

don't actually need the high resolution nor the whole brain devoted to get them done. So these parts would be taken out from those tweaks and ultimately one would achieve Ems still capable of operating what they were designed for, but they would be a lot cheaper because this process wouldn't need the whole brain to be successfully carried out. Hanson also relies on the fact that historically, as he claims, that's how most innovation has always been achieved: by trying out variations and that's what we'll do with brain emulations. (Hanson). Anders Sandberg also talks about the resemblance between my thought experiment and Robin Hanson's Em Economy, and highlights some of the differences between them, while also addressing some critiques to both of them although recognizing their hypothetical value. To him, after selecting my best me at a specific scenario in the kind of parallel evolutionary simulation involving the variation of hundreds of them, there would always be spurs left over, corresponding to uploads that would no longer have a future, which could be an ethical issue. "Hanson thinks that most people would be kind of happy to have a lot of spurs and delete the spurs." (Sandberg). But Sandberg does not think this to be true for most people although he hails the potentially powerful insight of making new modifications especially when we have opaque systems. And brain emulations are essentially black boxes. They are opaque systems.

If we can fully understand AGI, then we might get an alternative way to emulate, and Sandberg is convinced that this might be easier to improve than human brains because there is a backwards incompatibility between these two (AGI and human brain). He feels that combining human brains with AGI will be very powerful, but it is never going to be an easy task since we are dealing with quite different systems.

About the TE3b - iterating human intelligence

Em's economy is for Robin Hanson, as an economist, a natural download of mixing disruptive technologies with Darwinian principles. The uploads are, as he envisaged in his book and also here explained by him in the interview, emulations stemming out of specific regions of the human brain tied to specific functions, in order to also fulfill those specific jobs he talks about. Comparing this to my own TE, the following must be said: Hanson's model is not a WBE, rather partial emulations linked to their applicabilities. In the iterating human intelligence TE, a WBE is considered inasmuch as in order to iterate intelligence, its entire pattern of information and organization as a whole system would

have to be entailed. Hanson himself does not make a reference to intelligence, but in the Em economy it also would not make much sense to try developing specifically intelligence as a property, since intelligence does not really represent a function nor does it stand for a specific job, thus it would ultimately not necessarily lead to an increment in the Em's economy productivity. Furthermore, it would probably be insufficient to emulate one specific function to then postulate an increment of intelligence of the whole system. At most, it gets better in doing just that, but It does not seem probable that the intelligence of the original system, as a whole, could in this manner actually be augmented. It would very much depend on the context of that specific job (or generally speaking - function). One could also ask: how exactly would we identify the part that does the intelligence in that system? That is, which region, or sub-system would be responsible (if not the whole brain itself) for being intelligent, as if it could be regarded as a function of some kind.

Moreover, I would dare to argue that intelligence, pretty much like consciousness, is a property that cannot be that restricted, i.e., as discussed in the SI chapter (Part IV), it certainly involves more than just one aspect, function, mechanism or even property. If intelligence could be defined as the measurement of an agent's ability to achieve goals in a wide range of environments (Hutter, 2012), and if it involves a kind of optimization power / used resources, then I'm inclined to infer that to improve intelligence as a property, you would probably need to have the entire system working as a whole. Actually, is this not also one of the problems detractors of artificial intelligence usually charge it with, when dismissing its achievements? When Deep Blue beat Kasparov in chess in the 1990s, general criticism was that it was only good at playing chess, hence not necessarily intelligent in anything else (which by the way is actually true) and therefore not intelligent at all (in direct comparison to our human intelligence). To be intelligent means a lot more than just performing a specific task. AI refers to "systems which match or exceed human intelligence in virtually all domains of interest" (Shulman and Bostrom 2012). Thus, Ems' economy does not improve intelligence as a property, it improves specific tasks, to perform specific jobs. After all, that is the concept it was designed for in the context of economy.

In the TE-3b iterating human intelligence, we have an upload of ourselves, a whole brain emulation that is argued to improve in intelligence. One of the ways this could be achieved, would be by simple iteration of a millionfold of simulations in a specific scenario, like chess (to keep the example from above, and that actually was presented

to some of the interviewees). Now, today I might get to play 1000 games in my lifetime. Maybe more if I would have the time, but certainly in that same range or magnitude (assuming I would play twice a week I could end up playing 720 chess games in 10 years). So, I could never play out some $10E9$ games. It would take me millions of years, but I only have a couple of decades left (assuming everything goes well). Now, if I would play those 1000 chess games, it would naturally also mean I wouldn't be able to spend another 10 years playing Go. I would have to choose between spending time playing Go, improving in Go for that matter, or devoting myself to becoming a reasonably good chess player. Now, I am not going to argue that I am going to improve my intelligence simply by playing a lot of chess, though it would not be a bad argument for sure, especially when the order of magnitude would be in the $10E9$ range instead of those 1000. What I will argue is that I would be able to increase the mathematical combinations of all those scenarios. Not just chess, not just Go, but all possible mindsets there are. And all of them at the same time, since for my upload, time would not be an issue. My upload could have millions of copies. And still without altering anything to them (copies / uploads), still without any kind of intervention, adaptive or not, I could just run millions of uploads of mine, all at the same time, also - against millions of scenarios (besides chess, of course).

One of the advantages Deep Blue had (or any kind of chess playing computer for that matter) relied on the ability to preview millions of moves in about a second, forecast them all at the same time, and give an output result of the best possible move the next second. Imagine you would have all that. You could, with an upload. And because Deep Blue only knows how to play chess, whereas you (or your upload) know(s) lots of other stuff reasonably well (including playing chess), you would literally multiply your ability in solving all kinds of mind-puzzles.

Would you not improve your intelligence with the multitude of combinations available? And in a reasonable timeframe (actually in a few minutes, or seconds, as opposed to millions of years)?

Now, let us take even a step further. If we by then would be able to do a WBE, hence an upload of ourselves, then we definitely could also have that upload online. Meaning linked to the Cloud. Meaning, again, more combinations. Would this not improve even more intelligence? If we think of communication, as the interchange of information and if we consider this to accelerate the recombination of information, and if we consider the recombination of information to definitely be a catalyst in innovation, inasmuch as

recombining information results in creating new solutions, then we would have a positive feedback and we, too, could have a kind of biological (at least still, bio to some extent) recursive intelligence explosion.

Also, and taking one last step in the TE: Let us assume that corresponding to WBE we would get a digital file entailing the earlier mentioned pattern of information, which I called a kind of cognitive blueprint of ourselves, that would entail everything there is that defines us as ourselves. Today we have genetic sequencing which enables us to know our entire DNA sequence, specific to each and every individual. We can even identify and discriminate between different people based upon these exact genome sequences of ourselves. Is it a wild speculation to think that cognitive blueprint could be handled in a similar way DNA sequences are today? Even in that analogy with genetics of nowadays, one could speculate further and imagine a kind of editing of that cognitive blueprint in a very much similar way we do today with knock-in and knock-out genes. Now, I am sure we would be pushing limits in ethical terms in this latter step, but I am also quite confident that the former steps should not represent an ethical debauch.

David Pearce claimed that in order for us to have an exclusively biological superintelligence, we first have to solve the causal efficacy of consciousness. This might, of course, be true. However, as it seems, you could have intelligence without necessarily consciousness. Thus, we could develop or create superintelligence(s) without worrying with consciousness' hard (and easy) problems, as discussed in the SI chapter (Part IV of this thesis).

Table 16 – Selection of transcript excerpts regarding Question IV – Q2.

Question IV
Regarding Superintelligence and mind-uploading
<i>Q2: Could mind uploading prove that two different persons sharing an identical blueprint would in fact be identical? If so, would this not shake our beliefs in issues such as personal identity?</i>
RANDAL KOENE: “If their blueprints are completely identical, then they would have to be the same person. Even in twins it wouldn’t be the same. The only time that is going to happen is in fact at the moment where you make a copy of an upload and at that very moment before they start experiencing different things, they are in fact completely identical. But it means nothing because before they experienced things time hasn’t started to run for them, so they can’t sense anything, they can’t be aware of anything, they can’t feel that there is a problem.”
“Identity and consciousness, this thing that you can have in a state, where a state is something that you are, a specific point in time. But in reality, you’re talking about functions and things that produce an output over time. So, time has to go by to experience something to be aware of something to feel something. So only when these emulations are then progressing through time does the brain generate any experience.”
ANDERS SANDBERG: “Yes I think I’m with Parfit. This shows that personhood and personal identity isn’t a real thing. He was the guy who came up with the idea about personal identity in this Star Trek like tele-transport. He actually uses a bunch of examples where you merge people in various ways. So, you end up with two people with very similar or identical connectivities. I think Personal Identity doesn’t exist like we normally do think. Our concept of personal identity makes a lot of sense in everyday life just like free will makes a lot of sense on a human scale. I don’t think free will makes any sense on a neural scale or atomic scale, I think I’m a compatibilist, I think the

universe is essentially deterministic, give or take a quantum stuff, but that doesn't affect free will at all. Similarly, personal identity is kind of a macro thing happening when we talk about our lives and interact. But in a world where you can duplicate yourself or merge, people are fluid and, in that case personal identity stops being a powerful useful concept. We might need other more subtle concepts instead. That's going to be fun to develop, but it's probably going to be terribly annoying for a long while as we come to terms on how to do it. We're going to have something that makes sense for personal identity. It's just cultural evolution."

—

DAVID CHALMERS: "No, I don't see how it could prove this."

—

KENNETH HAYWORTH: "Every single one of these questions that involves personal identity and mind uploading and the philosophical copy problem, if you'd ask that same question to say C3PO, some robot with a computer brain, in that setting when you talk about C3POs computer brain copy happens to be identical to somebody etc etc ... If there's a simple answer from that second person, or third person view of a robot, and yet there's some mystery when it comes to human beings, you would reach rocket science, essentially."

—

NICHOLAS AGAR: "We'll need new concepts of identity! The digital intelligences will want things and value things. I care about a concept of identity that is within my physical self, but I bet they're not going to. Whatever concepts that we organize our values around won't be relevant or only be partially relevant to the concepts of those digital selves world."

—

RICK CATTELL: "If we could completely and accurately mind-upload, yes, we'd end up with two identical individuals until their life experiences caused them to diverge."

—

ROBIN HANSON: "When you have brain emulations of yourself you could make an exact copy of that emulation. That will almost surely have the same behavior as the initial until its experiences make it different then it becomes something with a different

behavior. So even if you and I have the exact brain state at this moment then we almost surely would behave exactly the same in the same situation. If you put us in different situations with different obstacles, then we would become different as a result of experience. The moment it makes a copy it's very close and predicts what the other would do very reliably and then they start having different experiences and start being more and more different."

RICHARD LOOSEMORE: "But they'd be different in the accumulated experience and knowledge. How much of a personality is the underline design and how much is the experience and knowledge? And we all know that the real answer is, almost all of it is in the experience and knowledge."

"There's no clear separation between mechanism and content, actually the content can change and influence the mechanism, and you can build new mechanisms in the same way that we can upload new software onto one hardware computer. This is obviously very flexible so there might be some flexibility and people might develop in thinking mechanisms that really are the mechanisms by which they think and it's not in the DNA, but in interactions between the DNA and the mechanisms the DNA design of that brain and its interaction with the world and as a result of the interaction of the world one person has a way of thinking about a certain problem that is so deeply embedded that is actually a mechanism something they use all the time and so deeply down there that they don't even know it's there."

"Still entertaining the mathematical possibility that even so, it just so happens that two people have the same mechanisms and the only difference is in the actual life experience and the content and the knowledge. I think it would mean very little I think the two of them would think in much the same way about any problem but that would only be the case as long as their past experience was not part of the solving problem."

Analysis #8:

Again, just like in the first question regarding personal identity (would a copy of me be me?), the psychological continuity, the experiencing through time, seems to make

the difference across the board to all interviewees, thereby refuting the mathematical and statistical probability of two equal beings existing at the same moment in time.

Randal Koene expresses and defines this cultural concept of identity and of consciousness as being somewhat as a state, a specific point in time whereas in reality they are just functions producing a specific output over time. The brain starts generating experience by progressing through time, so the upload and the original are only identical in the actual moment of making the copy, before time. For Sandberg this kind of experiment shows that personhood and personal identity is not a real thing and also that our concept of personal identity only makes sense in everyday life but as soon as we can duplicate ourselves, or merge; in this future world where people are more fluid, he claims this concept would disappear. He then foresees the need for a more subtle concept instead, which is also the reasoning made by Nick Agar. He foresees that uploads will value things in a different way than we do as biological humans. He highlights this with the fact that our identity concepts are tied to our physical selves whereas digital minds won't have this linkage we do.

Richard Loosemore, as basically all interviewees, underlines the accumulated experience and knowledge to be the difference that would separate the duplicates (and the copy from the original). Yet he questions how much of a personality is the underlying design and how much is the experience and knowledge? (Loosemore). He claims the answer to be obvious, which actually is in perfect alignment with all other interviewees as mentioned, but maybe because of a scientific and technological background as backbone of my own reasoning, I'm inclined to counter: Is there any way to measure it? No. Is there any empirical experiment undoubtedly proving that the experience of going through time would in fact make two identical people different? No. Yet there seems to be no doubt whatsoever.

Branching identity can provide specific predictions in even the most outlandish scenarios, as discussed in the MU chapter (Part II), and would in this case come close. (Cerullo, 2014). This identity theory alleges continuity of consciousness in multiple branches and can actually be applied to many of the exceptions occurring in nature such as the brain split syndrome. However, it is not the solution for this mathematical probability. I grant it to be a very low one and although feeling quite alone on this particular issue, I think that as long as there remains this non-zero probability, and also because none of the current identity theories can answer satisfyingly to naturally

occurring exceptions (such as split-syndrome cases) [§1](#) nor to the various uploading (or teleport) scenarios, and so identity remains an open question.

I will not try to argue there are two equal persons, let alone that we all are, or can be, to some extent the same person, as Hayworth expressed in his answer to the next question of this interview. In context, I do understand Hayworth's claim as a kind of wake-up call regarding an old-fashioned concept about identity we still have today. As Pearce used the expression of convenient myth in his interview, I, too, think that Identity theories are not scientifically sustained in order to satisfactorily solve the dilemma, which in my view, to some extent feeds that convenient myth. The importance of this directly relates to the fact that MU detractors use these identity theories (thereby, also the convenient myths along them) in order to evaluate and depreciate MU's eventual success in transferring our pattern of information that contains what we call consciousness and identity. Actually, detractors or skeptics of MU could better make their case in identifying identity-related problems that might emerge with MU, such as changing our brain's architecture. Some worry that enhancing intelligence may have a detrimental effect on personal identity. Agar suggests (elsewhere, not this interview, in "Why we should reject radical enhancement", Agar, 2013) that the procedures that might enhance our intellects will change the structure of our brains, thus leading to uncertainty about the identity of the person emerging from that procedure. Chalmers (also elsewhere, in "Mind Uploading: A philosophical analysis", Chalmers, 2010) also underlines the uncertainty regarding changing the brain's architecture. Extending intelligence indefinitely, and repeatedly, will surely carry a multitude of changes. And not only to our identities.

Table 17 – Selection of transcript excerpts regarding Question V – Q1.

Question V
Regarding Transcending Biology and Human Enhancements
<i>Q1: Is it inevitably going to generate a dystopian society?</i>
RANDAL KOENE: “I do believe it’s possible. I don’t often spend a lot of time thinking about that because as you may have noticed in the conversation, I really see the most important aspect of this being the adaptability that we gain and the way that we can then explore it into many possibilities as an advanced species.” — ANDERS SANDBERG: “I think you are assuming inequality is what gives you a dystopian society. But a lot of great literature dystopias have been about equal societies. The reason of course they are dystopias is not just inequality. And you can get all sorts of really weird disturbing dystopias you see in enhancements. For example, Nick Bostrom in his article “The Future of Humanity” (Bostrom, 2009), he brings up one where you have mindless outsourcers. So we outsource more and more of our brains to use for apps in the cloud and eventually there’s nothing left. So, you have an economy that is growing tremendously and super-skilled entities doing stuff but there’s no consciousness there’s nothing and nobody at home! He describes this as Disneyland without children. So, I don’t think it’s inevitable that we end up in dystopias. I think it’s inevitable that whatever the world turns out to be, somebody is going to say this is dystopia. But I actually don’t think that the world has necessarily become a dystopia that we all agree is a dystopia. I think we have a good chance of getting tremendously utopia worlds. Just if we think about the current world, it’s this mixture of horrific dystopia and beautiful utopia at the same time. All knowledge of mankind is more or less freely available to anybody and by the way we have oppression using facial-recognition in the interests of China it’s this crazy mixture of very good and bad and I expect this to continue so we will probably be horrified and delighted at the same time by the future whatever it turns out to be. Unless it goes

down in one of the straightforward dystopias.”

DAVID CHALMERS: “Not inevitably, but possibly. It depends on the AGI. (The AGI may be us.) Once we have uploading, relative immortality is a consequence.”

KENNETH HAYWORTH: “No, I’m definitely going to say no! The society at least understands how the brain works. And at a very deep level. Not today’s society where most people think they have immortal souls that are dropped from heaven or something and they really don’t understand this subtlety of uniqueness. In that world there is a different philosophy of morality. We haven’t seen it yet. But I think it will come about. In that world.”

“One of the perhaps logical conclusions of our really truly understanding the nature that we are, that we are computational, and our self-identity is really tied to a set of memories and not a point of view is to accept the idea that we’re all the same person. And that’s a philosophy that’s been around and I’ve written about it a little bit. But it’s a philosophy that I think will gain more and more traction as we become more comfortable with our scientific understanding of who we are and therefore I think society will become less and less dystopian. Because we will be adopting not a philosophy based on nothing, but a philosophy based on science that we are all the same individual. And I’d like to think that the dystopian possibility will become more and more remote because of that type of thing.”

RICK CATTELL: “Yes, all of this could lead to a dystopian society, and that’s scary because we only have one planet. What’s even scarier is that we don’t have “one planet” that is, if one country or one company does not develop an AGI, another one will. It’s like nuclear weapons in the 1950’s, or the super-intelligent computer in the movie Colossus. Ironically, the development of an AGI that rules the world, if we can design a benevolent one, could be the only way to save us from one that would destroy us!”

DAVID PEARCE: “Can one speak of a dystopian society in which everyone is blissfully happy? If one has other values beyond classical utilitarianism one may point and highlight features of our society that are sub-optimal. But it really is a civilization in which we have phased out the biological experience below hedonic zero, as will be feasible. We’re living in the last century of suffering that is biologically inevitable. Yes, you can’t have a dystopia in a traditional sense I would say.”

Analysis #9:

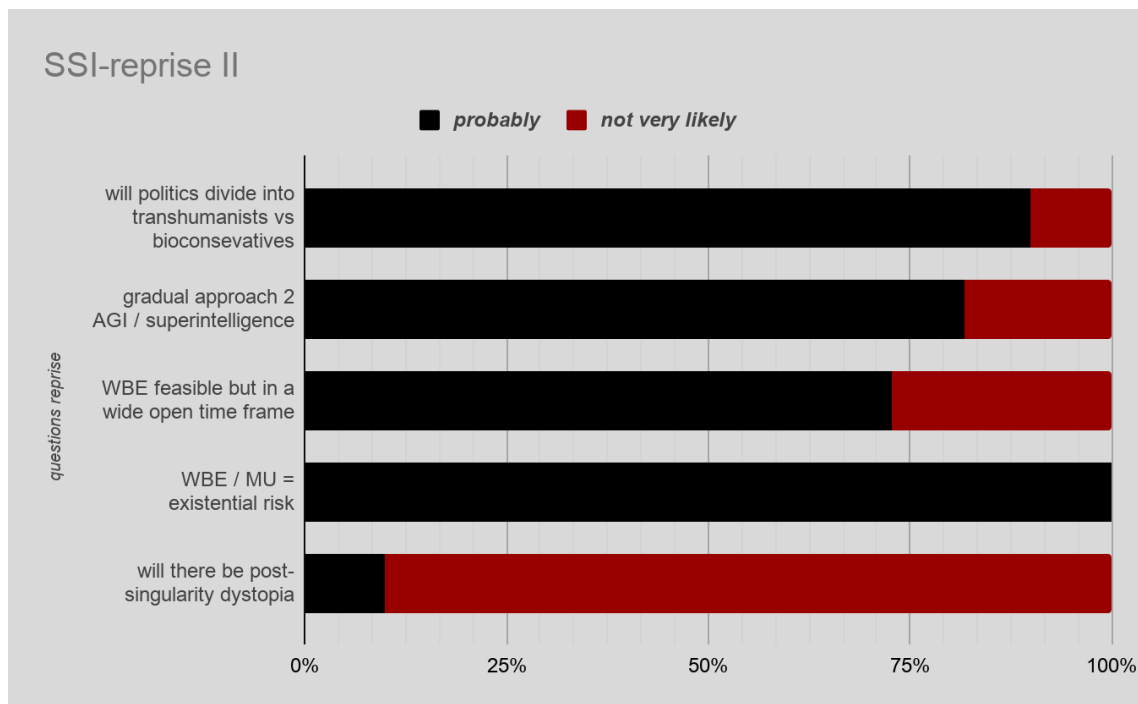
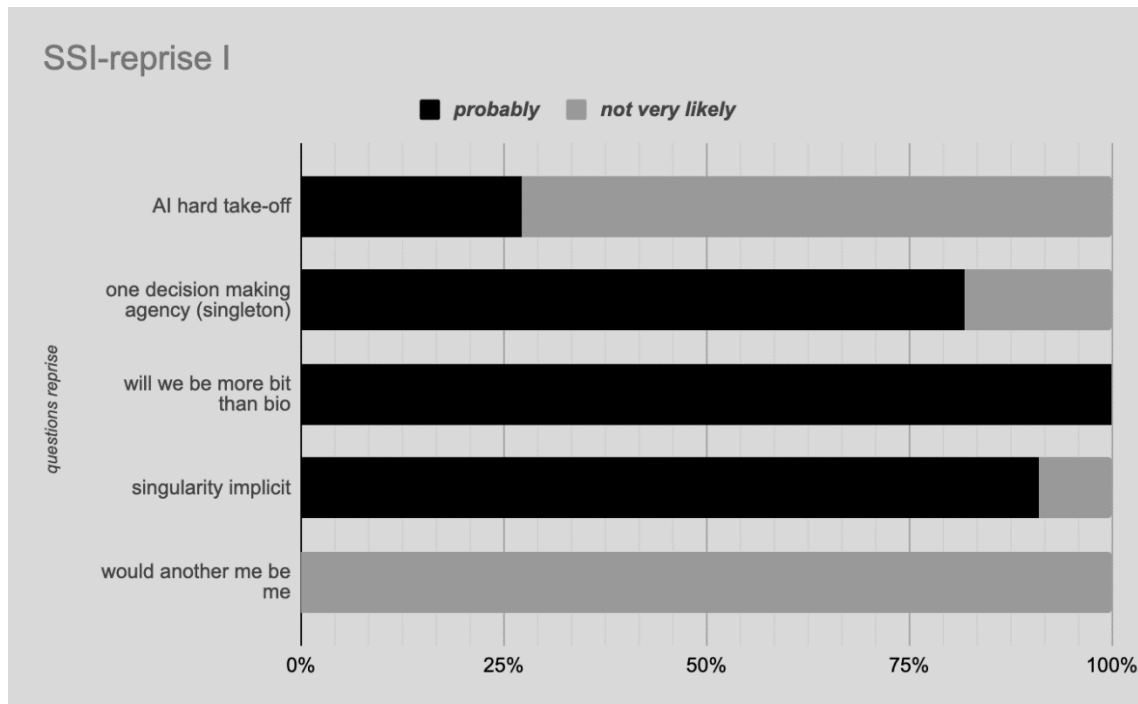
David Chalmers is actually one of the few who says a dystopia would not be inevitable, but possible and that this would basically depend on the kind of AGI. He also points out that the AGI might actually be us humans. Rick Cattell on the other hand thinks that it could turn out that ironically the development of a friendly AGI to rule the world could actually save us from one that would destroy us. Sandberg also thinks that whatever the world turns out somebody’s going to call it a dystopia, without it actually being one. He does rather believe in getting tremendously utopian worlds, instead.

Kenneth Hayworth strongly rejects the possibility of a dystopia. He again underlines that our self-identity is tied to a set of memories and not to a point of view. Furthermore, in a sense, he believes that “we’re all the same person” (Hayworth), to some extent, and that this philosophy, that is actually around for quite some time, will gain a lot of traction with us humans becoming more comfortable with our scientific understanding of who we are. He thinks that this can make society become less dystopian instead. He makes a reference to the fact that he has already written some things about this philosophy that is based on science and that through it we will end up acknowledging the idea that, as he claims: “we are all the same individual.” (Hayworth). David Pearce questions the likelihood of a dystopian society in which, as he claims, everyone would be blissfully happy. He evokes the fact that “we’ve phased out the biological experience below hedonic zero, as will be feasible. We’re living in the last century of suffering that is biologically inevitable.” (Pearce).

As a closing remark, once again, most of the authors seem to agree regarding dystopia. For them dystopia in a classical traditional sense, it is fair to claim, would not be a reasonable outcome. However, and although there are different possible definitions

on dystopia, depending on views and personal beliefs, it does seem undeniable that there might actually be two reasonably different kinds of humankind in this world of emulations, mind uploading and superintelligences. The bio-conservatism vs transhumanism clash that these interviews also addressed, may reach a peak by the time we would in fact have enhanced humans coexisting with non-enhanced ones. The differences would carry with them a whole set of other consequences eventually carving a deeper ditch in some kind of recursive feedback. The bigger the enhancements, the bigger the consequences, whether social, cultural and political: the loop, this recursive feedback of enhancements feeding consequences might literally divide our society in two inasmuch as by then, it would be difficult to argue not to live in a dystopia, or even argue to live in a multitude of utopias. Moreover, if these enhancements are to accelerate, i.e., are to be developed and ingrained almost in an exponential way, then the probability of coexistence of two quite different kinds of humans would increase in a direct proportionality. Thus, given specific outcomes of things such as the development of mind uploading or of superintelligence, one could speculate that the speed with which they might occur would actually be decisive in determining a dystopia or not. If these technologies are to be quickly ingrained and completely internalized, as already discussed in a previous question too, would we not witness a kind of shock? If, in the hard take-off scenario of an intelligence explosion, also already discussed, a kind of singleton were to emerge (as the majority of the interviewees hold to be quite possible, see SSI-reprise plot), then would we not inevitably have a dystopia?

As a conclusion, the following plots SSI-reprise I and SSI-reprise II show the interviewees' answers in a kind of snapshot and resume their standings in terms of probabilities regarding the issues discussed in the interviews.



Figures 30 & 31 – SSI-reprise I & II – show the pitch-kind of answers given by the interviewees to each specific issue in terms of probability. (Luig, 2020).

General Conclusions



Coming to the end of this thesis we need to recap some of its main claims. Each of the preceding chapters offered some preliminary concluding remarks, but now is the time to offer some general conclusions, grouped thematically and following the order of each chapter. Of these, special attention is given to the conclusions drawn from thought experiments because they involve all the other topics discussed in this thesis.

I am convinced that the fundamental ideas of science can be explained to anyone willing to make the effort needed to understand them. I am also persuaded that to understand these ideas is one of the most exciting and rewarding experiences one can have.

Thought Experiments

In this thesis, thought experiments are used as an instrument to discuss and communicate knowledge, following a logic of problem-posing / problem-solving and persuasion. In a first moment, they allow us to imagine possibilities that we didn't even know that existed. In a second moment, when we reflect on their implications, they can be useful to present possible solutions to the problems which arise through the scenarios they present. For instance, we learn of the existential threat that an AGI might represent to the survival of humankind through the thought experiments in which the possibility of an AGI arises; but then again, in my scenarios, the possibility of mind-uploading is a

possible solution to that threat – even though a solution that forces us to critically consider its (ethical and other) implications. Finally, and as I’ve argued throughout this thesis, thought experiments are a sound and efficient way to persuade – that is, to communicate – ideas, and mostly scientific ideas, because their narrative form is a simple and appealing way to convey arguments drawn from hypothetical scenarios. As such, in this thesis, the methodology of thought experiments boils down to the discussion and communication of science.

The thought experiments presented here were designed to fit the purpose and objectives of mind uploading and of superintelligence that are central themes of this thesis. Trying to conceive a TE that would somehow prove or demonstrate the probability of their feasibility or their occurrence, would have been artificially forced in the sense that: 1) it is not their supposedly higher probability rate that proves or actually demonstrates that they will occur and 2) there is already a large number of TEs that go precisely in that direction, which to me does not make much sense as can be understood from the above. No matter how well argued and rationalized they are, TEs will not and cannot prove something that has not yet occurred and that will only be verifiable through direct observance, empirical and scientific testing. Thus, the TEs presented in the thesis are used as a tool for “argumentative persuasion” of the feasibility and desirability of MU and superintelligence somewhere in the future. As such, we can say that the imagined scenarios of the thought experiments create a meaningful frame of discussion, suitable for the needs of peer-persuasion. When reading thought experiments, readers will be more easily convinced of the scientific possibilities that arise, and more compelled to draw conclusions and inferences from these possibilities. In the fields of ethics and political philosophy this is often the case, as exemplified by the thought experiments of the trolley problem or in Rawls’ veil of ignorance. But this is equally the case in science. And as such I believe that one of the outcomes of my thought experiments is to better convey the possibility of mind uploading and superintelligence – and through them I am effectively communicating science – and also of one of its main implications, that of the survival of humankind through different means if faced with an AGI.

To reinvent identity

The TE 3a) *iterating human predictor response* argues about the possibility that MU can fine-tune the cut-off of possible actions and behaviours we commonly have, in such a way as to make possible for us to forecast many of our actions, responses, behaviours and ultimately even thoughts, i.e., to make them predictable. In its last scenario, the TE goes to the limit of actually considering the hypothesis that the shortening of that cut-off window ultimately narrows people's differences to a point in which it becomes difficult to distinguish their identities inasmuch as their overall behaviour somehow reflects their personal idiosyncrasies. As argued by some of the experts interviewed for this thesis, MU might show us that personal identity theories will have to be revised and this specific last scenario of the TE supports precisely that idea.

Generally speaking, scientific and technological progress has been going beyond borders and narrowing differences, sometimes even blurring formerly well-established concepts. Many of these conceived of the human being as something unique and special. However, progress has been withdrawing some of that uniqueness, to the point of ultimately even making us something trivial. But that is not actually true if we regard the whole picture instead of focusing on details, i.e., if we focus on the fact that MU might lead us to *reinvent identity* this should not diminish our humanity, rather shift its course.

To keep the competitive balance

The TE 3b) *iterating human intelligence* delves on the possibility of combining millionfold scenarios with millionfold copies; i.e., it argues that while it would be physically impossible for us today to increase the combination of mathematical variables that have the power to exponentiate our intelligence, yet this might be feasible through MU. In fact, a multitude of different scenarios we might be exposed to at the same time would definitely increase our capacities and abilities in previewing, forecasting, evaluating reflectively etc. Furthermore, and considering the possibility of being linked together in *the Cloud*, this would also foster the recombination of information, which might then even allow a kind of recursive intelligence explosion in a way pretty much similar to what we today envisage as being AI's eventual future development. It is also argued through this particular TE that in order for us humans to maintain a competitive balance with that

probable AGI upcoming development, we have to bet on our own intelligence's enhancement. Or, if not, face a quite probable scenario of being outpaced with still unpredictable outcomes, including a potential existential risk.

We can blend with the technology where and when it makes sense to keep our fate as a species in our own hands, if the following conditions are met: 1) assuming our intelligence to be the key-tool that allows us to until some extent control the course of our own evolution; and 2) also assuming a slow kind of AI take-off, hence a singularity in a wide period of time, also as hypothesized by the majority of the experts interviewed for this thesis; and 3) controlling AI's development by ingraining it with human values, or at least developing it inasmuch as to allow a friendly AI to emerge, then iterating our own intelligence via MU could take us to another competitive and adaptive level. The TE ultimately hypothesizes about our chances of self-improving in a similar way AI does, but of course only through our uploads (SIMs) and thus maintain flexibility to adapt to new environments and new circumstances.

Some advantages might be grasped in this scenario. On the one hand, there is no biological pandemic that can wipe out software. On the other hand, of course, digital disasters are also easily imaginable: there could be an outbreak of computer viruses wiping out brain emulations! However, this might be another argument for developing MU and SIMs with augmented competitive intelligence: although brain emulations could be vulnerable to computer viruses, they would also be a backup of biological humans whereas they, who are in turn vulnerable to biological viruses, would also in some way be a sort of backup of their uploaded counterparts. Hence, by splitting humankind in two fronts so to speak, the overall risk would be significantly reduced. Furthermore, *superintelligent uploads could bypass AGI's risks by replacing a flawed self-improving machine intelligence (e.g. with flawed motivation systems and dubious goals) with copies of reasonably well understood human intelligence*. So, in one scenario AGI could generate rapidly upgradable software minds without human values and in the other our human uploads would slow the artificial superintelligence hard take-off and allow for a more insightful control both over the future of *our own kin* and also of that software: "Unlike de novo artificial intelligence it appears possible to benchmark progress towards brain emulation, promising a more concrete (if arduous) path towards software intelligence". (Sandberg, 2014, p. 35).

To perpetuate consciousness

The TE 3c) *the delusion of reality* argues for the necessity of perpetuating consciousness giving meaning to life and existence altogether. This TE explores the metaphysical purpose of MU along with the possibility of an upgrade of our consciousness level, assuming it would follow an intelligence's upgrade. AI, or AGI, or even an ASI, would probably not perpetuate humankind's consciousness nor humans' personal identities. But MU could. It presents itself as the only envisaged technique or technology allowing preserving and therefore also perpetuating human consciousness.

Consciousness may have developed over millions of years and let us assume it may even have had its origin far back having originated in behaviour and in the predecessors of detection, learning and memory. It is argued that the complexation of brains and of nervous systems enabled capacities such as prediction, planning, remembering through storing of patterns of information and learning, ultimately allowing the regulation of behavioural responses and thus, also the evaluation of possible outcomes of behaviour. Thus, consciousness might have evolved as a way of going beyond merely responding reflexively to stimuli, i.e., it might have developed so as to allow more flexible and more planned responses, and so permitting a balanced decision on the overall response or behaviour to adopt before actually doing it. In this evolutionary course, this might have ultimately led to that basic sense of self in complex organisms such as mammals and even to the ability to better represent internally thinking patterns more reflectively in primates and especially in humans, i.e. to be aware of being aware. (Feinberg and Mallatt, 2016).

Feinberg and Mallatt argue that a primary consciousness (also called sensory consciousness, or simply the most basic form of consciousness), might have originated in the first vertebrates and then through various evolutionary paths got increasingly complex making the emergence of subjective experience possible, i.e., the emergence of being aware of stimuli and of the environment, but with no reflection at all. This would correspond to feeling "something that it is like to be", as Nagel phrased it in his famous essay "What is it like to be a bat?" Nervous systems kept on developing and eventually neurons became clustered together in networks allowing, among other things, to store and recall experience, thus making learning and memory available. Hence, the processing of information became ever more complex, and way more complex than

simple reflexes as answers. With more elaborately connected brains, also sensory consciousness became further enriched with internal representations not only of the environment but also of the self - a kind of mapping in the brain of the body and actions. (Feinberg and Mallatt, 2016).

So, if we accept 1) that intelligence and consciousness may have developed arm-in-arm; 2) that both of them increment our ability of understanding, interpreting and even predicting and forecasting our environment, and 3) if MU can augment intelligence exponentially, then it might therefore also take consciousness to an upgraded level. Also, if, as hypothesized in this TE, virtual reality might be augmented so as to cast doubt about the authenticity of reality in an uploaded and upgraded mind, then, that this leads to a search of meaning for life and existence altogether does seem like a reasonable inference to make, in this non-zero probability scenario. Might this superconsciousness state, or collective consciousness state, or even this matrix not be representative of human thought? And would not this human thought, even though a product of MU be worth preserving and perpetuating? In my view, it would definitely account as a backup of humankind against possible extinction scenarios whether of anthropogenic nature or not. Hence, MU might also work as a survival backup as opposed to an existential risk. Furthermore, with an upgraded superconsciousness and superintelligence, humankind might not just find meaning in life and existence as mentioned, but even give meaning to that existence in this universe, or conversely give this universe a meaning by finding a purpose for that *consciousness 2.0*.

WBE: The Feasibility Discussion

WBE may be trending to be *the technology of preserving, scanning, imaging, mapping and functionally reproducing the brain*. The main concern of the WBE chapter revolves over the technical discussion of its feasibility. As the plots A and B (Figures 06 & 07) of that chapter's analysis clearly show, there are some difficult barriers to transpose in order to achieve a WBE of a human brain in the coming years. However, it is also argued in that chapter, that they are transposable and that the real discussion makes more sense to be held over its timeline, than actually over its occurrence or not.

MU: Reinventing identity, hence reinventing consciousness (or the other way around)

MU, if feasible, might lead to a reform of concepts such as personal identity, and even of consciousness too. In fact, many of the issues discussed in this chapter are controversial. Many are open questions and even many more of them are quite speculative. Nevertheless, the chapter ends with a couple of concluding remarks: 1) consciousness is not human-exclusive and may be generated in other than biological substrates and 2) consciousness may be measurable, and so we may be able to compare different sentient entities and therefore also develop a criterion for validating a mind uploading. Physicalists, functionalists and patternists consider MU to become feasible through brain preservation and later on through destructive uploading, therefore enabling the preservation and perpetuation of consciousness. The metaphysical discussion of this purpose under the light of *existentialism*, such as we attempted in that chapter, allows us to explore consciousness on several levels and upon various views and exponentiates MU's potential in disclosing that particular debate.

Even though one can argue that it still is speculative whether MU will in fact ever take place or not, its non-zero probability and the large amount of possible outcomes and in particular, outcomes that might be disruptive for humankind, makes it desirable to discuss it in advance so as to be better prepared for what might happen. The emergence of superintelligence is such a possible scenario worth discussing and preparing for.

SI: The difficulty and fear of dealing with the unknown

Independently of the kind of take-off that AI will most probably have, there won't be much time for humankind to react once an intelligence explosion has been onset. Thus, as mentioned above in this conclusion, the TE-3b, *iterating human intelligence*, might work as a strategy for humankind to maintain a competitive balance with AI and along with other measures such as developing friendly AI, prevent an unsafe development or course of events due to exponential self-improving AI. This chapter discussed several approaches to deal with an emergent non-human powerful intelligence and exposed conflicting views in current strategies to control this development. Ultimately there seems to be no consensus today about whether this development would be better safe-assured

as an open-source kind of development or on the other hand confined to a restricted group of people. This disagreement in ideas and opinions very much reflects the actual status quo in not knowing how to deal with the unknown, which is also a fertile ground for settling the clash in coming years.

It is also argued in this chapter that AI might have to require self-awareness in order to determine its own goals. Hence, a *singleton* might entail a conscious state and therefore constitute a non-human consciousness empowered through a strong intelligence. I have therefore speculated towards the end of this chapter about the possibility of our fear of such an entity and, based on our anthropomorphic bias, about the fear of such an entity's consciousness resembling a human exponentiated kind of dangerous consciousness. Finally, a singularity driven by uploads instead of by AI, although less probable according to the majority of the experts on this particular view, would by default carry a lower risk associated because of our human participation, and thus a desirable kind of outcome.

SSI-Analysis: Some uniformity among the interviewed experts

From a direct and simple observation of both SSI-reprise plots (Figure 30 and 31) made from the answers given by the experts interviewed for this thesis, it is possible to confirm, in that summary, that in terms of the probability and likelihood of occurrence of the various issues discussed, some degree of consensus among these experts was patent. There is actually not a single issue that really divided them deeply enough for us to speak of a real controversy. Either all of them or at least the majority always inclined towards one possible outcome. Each issue / outcome was either considered to be probable or on the other hand not very likely in an almost unanimous way. The fact that the majority of the answers have not been scattered throughout several possibilities provides some consensus, a general feeling that might be a reflection of a common sense kind of knowledge, not only shared by experts, but also by non-experts as the plot (Figure 29) of the SI chapter showed concerning prediction of AI-related issues.

Limitations and recommendations for future research

Arriving at the end of this doctoral thesis, some words must be said on the specific limitations of the research that has been put into it. First, some of them are connected with the choice of the topic itself. As should by now be clear to the readers, whole brain emulation, mind uploading superintelligence are “only” hypotheses, and working on them, assessing the likelihood of predictions on future technological developments and so forth involves a good deal of speculation, even if this speculation is based on solid knowledge and cogent arguments. As such, in this thesis I am not communicating and discussing past science (this is not a work in the history of science) or, for the most part, even present-day science. On the contrary, and to some extent, what I am discussing and communicating is the *future* of science and one of the most important – perhaps *the* most important? – ways in which it will impact the future of humankind. But the fact that this relies mostly on hypotheses can of course be seen, somehow, as a limitation, because there will always be ‘naysayers’. I have done my best to assess these possibilities in a critically informed way, but it goes without saying that some of the claims I put forward here might be proved wrong, because the future is always somehow unpredictable, at least in some of its aspects.

A second limitation has to do with the context in which this thesis has been put forward. As I noted at the outset, the thesis deals with topics that are almost totally unexplored in the Portuguese scientific and intellectual landscape. As such, and in this context, they come with a ring of novelty and this is not always the best context for their reception. But science has always had to do with disruption; therefore, even though this might be a limitation, advancing – and discussing – disruptive or polemic ideas is also something that, to my mind, the communication of science needs to put forward.

A third limitation involves the methodology of this thesis. As explained in the introduction, this is a highly interdisciplinary endeavor with contributions and methodologies that range from biology the neurosciences and NBIC technologies to philosophy and the social sciences, all of them combined to put forward an effort in the discussion and communication of science. To my mind, this is a strength of this thesis but, from other angles, it can also be seen as a limitation given that within each specific field other hypotheses, interpretations or methodologies might be proposed. But given that I focused on problems and possibilities rather than on accepted dogma, it might be said

that only such an interdisciplinary approach could – even though with all its flaws – begin to grasp the complexity of these problems.

Also, a limitation that follows from this methodology, and that might eventually arise as an objection to my approach to these subjects, is that I did not devise a new methodology for the communication of science *strictu sensu*, i.e. I did not directly engage with the public in any way whatsoever during the course of the thesis trying to communicate science. This is true, and it could be considered a limitation. However, in what concerns this point I must reiterate that I see the work done for this thesis as a starting point, rather than an end point for my research in these matters. As I have tried to argue throughout the thesis, my methodology to communicate science was based in the tightly argued critical assessment of these topics, in the design of the thought experiments, and in the interviews conducted with experts. I believe that the thought experiments I designed can and indeed will serve – if used – to better communicate these burning scientific topics so important for the future of humankind.

In what concerns the interviews my plan is to come back to them and to use them to reach a broader audience, thus taking my effort to communicate and discuss science a step further. As I explained in the SSI analysis section of this thesis, when the interviewees were contacted their consent was explicitly granted to use this material solely for the academic purposes of this thesis. But the way the interviews were conducted – with the relative freedom that a format of semi-structured interviews allows – and the richness of their results made me reflect on what possibilities laid ahead. As such, my plan for the immediate future of this research is to again contact the authors and to revise the interviews – allowing the authors to co-edit them with me for purposes of clarity and scope of content – in order to publish them together in book format, which would thus be yet another output of this research, and one with clear results for the communication of science.

Another limitation also connected with the interviews has to do with their replicability. It might be argued that their script was devised with questions so tailor-made for experts that it might be hard to replicate these interviews in other contexts. Granted, this problem will evidently arise should one want to replicate them with non-experts. But then again, the guide, as it is, is perfectly feasible to apply with other experts, even though with a caveat: given that my own thought experiments are discussed, the interview evidently makes more sense if being conducted by me. As such, replicability was not an aim in

itself but another possibility for the future of this research is simply to address other experts and conduct interviews with the same guide.

Finally, some recommendations for future research can be spelled out. And here I am not primarily referring to my own future research, whose avenues for the development of these issues I already sufficiently sketched, but rather for others' research. As stated, my general aim at this point of my research was not to directly be in touch with the general public but it goes without saying that that dimension is very important for the communication of science. As such the degree of acceptance of topics such as mind uploading could be further advanced by using some of the outcomes of this thesis, such as the scenarios of the thought experiments, and their implications. As such, future research could use what I take to be some of my contributions to the field, such as the thought experiments.

It would also be interesting to further investigate the conditions that would have to be met for people to willingly chose to undergo mind uploading. Such a research is non-existent in the Portuguese context and it would not be futile to attempt it. As regards the teaching of science, even though this is not the specialization of my thesis, I would argue that the topics approached in this work could very well, in the future, be present in the curricula given their importance for our future. Because it is better to anticipate future possibilities than to simply have to deal with them when they do come about.

Final note (Ethical Disclaimer)

As mentioned in the Introduction, the thought experiments explore the possibility of manipulating million-fold copies of oneself. In the event, as argued in this thesis, that those uploads might be sentient beings, we would have crossed ethical boundaries as mentioned on several occasions throughout the dissertation. Nevertheless, and still without knowing exactly what this would mean in practical terms to those specific uploads, we have to approach this issue from the overall perspective of humankind, in the sense that if in direct competition with an eventual ASI, our human species' survival might depend upon MU. This can perhaps be approached more or less as the trolley problem, insofar as a decision with ethical implications might have to be carried out sometime, with the outcome never being perfect, but forcing us to chose nonetheless.

This is actually a pertinent question since AI, or at least the ongoing artificial narrow intelligence would probably choose the lesser impactful route even if this one would entail some collateral damage. As such, and as I argued in the Ethics chapter, even though this TE can certainly go against the grain of some ethical intuitions that some people might have, given that we would be dealing with an extreme scenario, perhaps a utilitarian judgment would apply. In that case, pitting the manipulation of the emulations against the eventual demise of humankind, perhaps that manipulation (even though ethically problematic) might be the lesser of two evils.

If we choose intelligence as our main asset in competitive terms, then we definitely have to improve it in order to cope with artificial intelligence or face the true possibility of extinction. However, and as argued in the MU chapter when exposing these TEs, any kind of scenarios that might be considered as “harmful” could with the right kind of algorithm applied be avoidable, thus only exploring *consensual, or at least (arguably) ethically defensible scenarios*. We might even envisage an ethical committee supervising the nature of these scenarios and still, through another million-fold of other harmless scenarios in a controlled environment explore the full potential of these TEs. The underlying argument would be the fact that ultimately humankind’s perspective should prevail over the interests of the individual. And in this, MU might prove useful.

I have tried to concentrate on problems rather than on accepted truths, but this at once raises a difficulty. It is impossible to discuss the questions we cannot answer without first answering those that we can. My contribution was partially to show the scenarios under which these questions arise. Whether or not I was successful in communicating these topics and the main implications they involve, it will be up to the readers to judge.

References

- Agar, N. (2011). On the irrationality of mind-uploading: a reply to Neil Levy Nicholas, Ó Springer-Verlag London Limited 2011, AI & Soc (2012) 27:431–436.
- Ajay and Bhalla, (2006). Synaptic plasticity in vitro and in silico: Insights into an intracellular signaling maze. *Physiology*, 21: 289-296, 2006.
- Albantakis L, Hintze A, Koch C, Adami C, Tononi G (2013). Information Matching – Environment dependent increase in integrated information (W). European Conference on Complex Systems (ECCS13).
- Alivisatos, A. et al (2012). The Brain Activity Map Project and the Challenge of Functional Connectomics. *Neuron*, 74: 970 – 974.
- Alivisatos, A. et al (2013). The Brain Activity Map. *Science*, 339: 1284 – 1285.
- Allen, P., & Greaves, M. (2011). The singularity isn't near. *Technology Review*. October 12, 2011, <http://www.technologyreview.com/blog/guest/27206>.
- Alsheheri, M (2012). Interaction of double-stranded DNA inside single-walled carbon nanotubes, Springer Science.
- Alsheheri, M (2014). Modell Interaction DNA with carbon nanostructures, PhD Thesis in Applied Mathematics, University of Adelaide.
- Armstrong, S., K. Sotala and S. S. Ó hÉigearthaigh (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. *Journal of Experimental & Theoretical Artificial Intelligence*.
- Ashby, F. Gregory, and Sebastien Helie. (2011). A Tutorial on Computational Cognitive Neuroscience: Modeling the Neurodynamics of Cognition. *Journal of Mathematical Psychology* 55 (4): 273–289.
- Bainbridge, W. (2012). Whole personality emulation. *International Journal of Machine Consciousness*, 4 (1), pp. 159–175.
- Bainbridge, W. (2006). Strategies for personality transfer. *Journal of Personal Cyberconsciousness* 1(4), 13!14,

<http://www.terasemjournal.org/PCJournal/PC0104/bainbridge01b.html>.

Bainbridge, William Sims, and Mihail C. Roco, eds. (2006). *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society*. Dordrecht, The Netherlands: Springer.

Baum et al (2017). A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis, *Journal of Experimental & Theoretical Artificial Intelligence* 29(2) 397-414.

Baum, S (2018). *Superintelligence Skepticism as a Political Tool*, Information 2018, 9, 209; doi:10.3390/info9090209.

Bamford, S. & Danhaer, J. (2017). *Transfer of Personality to a Synthetic Human ('Mind Uploading') and the Social Construction of Identity*, Journal of Consciousness Studies, 24, No. 11–12, 2017, pp. ?–?

Bardziński, F (2015). *Between Bioconservatism and transhumanism: In search of a third way*, Ethics in Progress, Washington.

Barrat, J. (2013). *Our Final Invention (Artificial Intelligence and The End of The Human Era)*, St Martin's Press.

Bensaude-Vincent (2009). *Les vertiges de la technoscience: façonner le monde atome par atome* La Découverte, Paris.

Berger, et al (2012). A hippocampal cognitive prosthesis: multi-input, multi-output nonlinear modeling and VLSI implementation. *IEEE Trans Neural Syst Rehabil Eng* 20(2):198–211.

Berger, et al, (2011). A cortical neural prosthesis for restoring and enhancing memory, *Journal of Neural Engineering* 8, 4, August 2011.

Berger, H. (1929). *Ueber das Elektroenkephalogramm des Menschen*. Archiv fuer Psychiatrie und Nervenkrankheiten, Berlin 87, 527–570

Berger, T., and Glanzman, D. (eds.) [2005]. *Toward Replacement Parts for the Brain: Implantable Biomimetic Electronics as Neural Prostheses*, (MIT Press).

- Berger, T.W., et al. [2005]. Restoring Lost Cognitive Function. Hippocampal–Cortical Neural Prostheses. IEEE, *Engineering in Medicine and Biology Magazine* 24(5), 30-44.
- Blackford, R. and Broderick, D. (2014). *Intelligence Unbound: The Future of Uploads and Machine Minds*, Wiley Blackwell, Chichester.
- Blaschke, B. (2017). Mapping brain activity with flexible graphene micro-transistors. *2D Materials*, Volume 4, Issue 2, article id. 025040 (2017).
- Blattner, W. (2017). Anonymity, Mineness, and Agent Specificity: Pragmatic Normativity and the Authentic Situation in Heidegger's Being and Time. In *Giving a Damn: Essays in Dialogue with John Haugeland*, edited by Zed Adams and Jacob Browning (Cambridge: MIT Press, 2017): 51-72.
- Block, N. (1995). On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, 18(2), pp. 227–287. BMC Neurosci. 5, 42.
- Bokulich and Frappier (2017). *On the Identity of Thought Experiments: Thought Experiments Rethought*. The Routledge Companion to Thought Experiments.
- Bolzan, S., Bolzan, J. & Rother, C. (2011). *My Life, Deleted: A Memoir*, San Francisco, CA: HarperOne.
- Bonilla, J. Z. (2017). *Your mind will never be uploaded to a computer*. Mapping Ignorance. <https://mappingignorance.org/2017/09/11/mind-will-never-uploaded-computer/>
- Bostrom, N. (2002). Existential risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9.
- Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in *Advanced Artificial Agents, Minds and Machines*, Vol. 22, Issue 2.
- Bostrom, N. (2013). Existential risk prevention as global priority, *Global Policy* 4, 15-31.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.
- Bostrom, N. (1998). How long before superintelligence? *International Journal of Future Studies* 2.

- Bostrom, N. (2005). History of Transhumanist Thought. *Journal of Evolution and Technology*, 14, 1–25.
- Bostrom, N. and Savulescu, J. (2009). *Human Enhancement*, Oxford University Press, Oxford.
- Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211), pp. 243-255.
- Brown, J. (1991). *Laboratory of the mind: Thought experiments in the natural sciences*. London: Routledge.
- Brown, J. et al (2018). *The Routledge Companion to Thought Experiments*, Routledge.
- Brueckner, A. (1992). Conceiving one's envatment while denying metaphysical realism. *Australasian Journal of Philosophy*. Vol 70, No. 4.
- Calafate, L. C. (2019). A Experiência Mental no Ensino das Bases Neurobiológicas da Tomada de Decisão Moral: O Dilema do Elétrico. *XVIII ENEC / III SSE, Educação em ciências: cruzar caminhos, unir saberes*, Book of Abstracts, ed. by Vasconcelos, C. Ferreira, R. A., Calheiros, C., Cardoso, A. Mota, B. & Ribeiro, T. Porto: Porto University Press, pp. 377-378.
- Camus, A. 1961. *The myth of Sisyphus and other essays*. Vintage Books.
- Cappuccio, M. (2017). *Mind-upload The ultimate challenge to the embodied mind theory*, 16(425-448) Phenom Cogn Sci.
- Carvalho Guerra, I. (2006). *Pesquisa qualitativa e análise de conteúdo – sentido e formas de uso*. Princípia Editora, Estoril.
- Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine* 5 (198): 198ra105–198ra105.
- Cattel, R & Parker, A (2012). *Why is Building a Brain so Difficult?*, SynapticLink.org & Department of Electrical Engineering University of Southern California.
- Cerullo, M. (2015). Uploading and branching identity, *Minds and Machines*, Springer Netherlands, 25(1), pp. 17–36.

- Chaisson, E. (2012). *A Singular Universe of Many Singularities: Cultural Evolution in a Cosmic Context*, in A. H. Eden et al. (eds.), *Singularity Hypotheses*, The Frontiers Collection, 413 DOI: 10.1007/978-3-642-32560-1_20, Ó Springer-Verlag Berlin Heidelberg 2012
- Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press, Oxford.
- Chalmers, D. (2010). *The Singularity, A Philosophical Analysis*, Journal of Consciousness Studies, Volume 17.
- Chalmers, D. (2014). Uploading: a philosophical analysis, in *Blackford and Broderick* pp 102- 118.
- Chandrasekharan, S. & Nersessian, N. (2015). *Building Cognition: The Construction of Computational Representations for Scientific Discovery*. Cognitive Science, Wiley.
- Cho, Y (2012). *Polymer and nano-technology applications for repair and reconstruction of the central nervous system*, Journal of Experimental Neurology.
- Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.
- Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis* 58:7-19.
- Cooper, R. (2005). *Thought Experiments*. Metaphilosophy LLC and Blackwell Publishing Ltd.
- Corabi, J. (2014). *The Metaphysics of Uploading*. Uploaded Minds, Russell Blackford, Wiley-Blackwell.
- Corabi, J. & Schneider, S. (2012). The metaphysics of uploading, *Journal of Consciousness Studies*, 19(7–8), pp. 26–44.
- Cordio, M. (2008). Searle's Chinese Room Argument and its Replies: A Constructive Re-Warming and the Future of Artificial Intelligence, *Indiana Undergraduate Journal of Cognitive Science* 3, pp 51-57 Copyright © 2008 IUJCS.
- Crevier, D. (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books.
- Crick, F. Koch, C. (2003). A framework for consciousness. *Nat Neurosci* 6: 119–126.

CyBeRev: <https://cyberev.org/>

Dawkins, R. (1976). *The Selfish Gene*. New York City: Oxford University Press.

Dawkins, R. (2000). *The Blind Watchmaker*. Penguin Press Science.

Deca, D (2011). *Available Tools for Whole Brain Emulation*, International Journal of Machine Consciousness.

Del Monte, L. (2013). *The Artificial Intelligence Revolution (Will Artificial Intelligence serve us or replace us?)*.

Dennett, D. (1991). *Consciousness Explained*. Little, Brown and Company.

Dennett, D. (1991). *Granny's Campaign for Safe Science* In: B. Loewer and G. Rey (eds.) *Meaning in Mind: Fodor and his Critics*, Oxford: Blackwell.

Dennett, D. (2003). Explaining the Magic of Consciousness, *Journal of Cultural and Evolutionary Psychology*, 1(2003)1, 7–19

Dennett, D. (2005). *Darwin's Dangerous Idea*, Touchstone Press, New York pp. 352 to 360.

Dennett, D. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*, W.W. Norton & Co, New York.

Dennett, D. (1988). Quining qualia. In *Marcel & Bisiach (Eds.), Consciousness in contemporary science* (p. 42–77). Clarendon Press/Oxford University Press.

Denzin, N. K. & Lincoln, Y. S. (eds.) (2018). *The Sage handbook of qualitative research. Fifth edition*. Sage, Los Angeles.

DeRose, K. (2000). How can we know that we're not brains in a vat? *The Southern Journal of Philosophy* (2000) Vol XXXVIII.

Di Paolo, E. Noble, J., Bullock, S. (2000). *Simulation Models as Opaque Thought Experiments*, - [apophenia.wdfiles.com /](http://apophenia.wdfiles.com/)

<https://eprints.soton.ac.uk/261455/1/opaque.pdf>

Dorrier, J. (Jan 2015). *If you upload your mind to a computer—are you still you?* Singularity Hub, Hill, D. (ed.).

<http://singularityhub.com/2015/01/25/if-you-upload-your-mind-to-a-computer-are-you-still-you>

- Dreyfus H. (1965). *Alchemy and AI*, RAND Corporation, Santa Monica, California.
- Dreyfus H. (1997). *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press, Cambridge, Massachusetts.
- Dreyfus, H. (1972). *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row.
- Drexler, E. (2013). *Radical Abundance (How a Revolution in Nanotechnology will change Civilization)*, Public Affairs, New York.
- Ebbesen, M (2006). *Ethics in Nanotechnology: Starting From Scratch?*, Bulletin of Science, Technology & Society.
- Edelman, G.M. (2003). Naturalizing consciousness: a theoretical framework. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5520–5524
- Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. (2012). *Singularity Hypotheses: A Scientific and Philosophical Assessment*. The Frontiers Collection. Berlin: Springer.
- Elgin, C. (2014). Fiction as Thought Experiment. *Perspectives on Science* 22(2), pp221-241.
- Elgin, C. (2017). *True Enough*. MIT Press. Cambridge, Massachusetts.
- Eth, D., Foust, J., & Whale, B. (2013). The prospects of whole brain emulation within the next half-century. *Journal of Artificial General Intelligence*, 4(3), 130–152.
- Feierbach & Koene, (2015). *Whole Brain Emulation - State of the Art*, https://www.researchgate.net/publication/290820977_Whole_Brain_Emulation_-_State_of_the_Art
- Feinberg, T. and Mallatt, J. (2016). *The Ancient Origins of Consciousness: How the Brain Created Experience*. Cambridge, MA: MIT Press.
- Feldman, J. and Ballard, D. (1982). Connectionist Models and Their Properties. *Cognitive Science* 6 (3): 205–254.

- Fields, (2013). Map the other Brain, *Nature*, 5; 501 (7465): 25-27.
- Foot, P. (1978). *Virtues and Vices and Other Essays in Moral Philosophy*. Oxford University Press.
- Fortin, M-F. (2009). *Fundamentos e Etapas do Processo de Investigação*. Lusodidacta, Loures.
- Freitas, R. (1999). *Nanomedicine, Vol. I: Basic Capabilities*, Landes Bioscience, Georgetown, TX.
- Freitas, R (2009). *Meeting the Challenge of Building Diamondoid Medical Nanorobots*, Institute for Molecular Manufacturing, Palo Alto, CA, USA.
- Fukuyama, F. (2002). *Our Posthuman Future: Consequences of the Biotechnology Revolution*, Farrar, Straus and Giroux, New York.
- Fuller & Lipinska (2014). *Transhumanism*, Social Epistemology Review and Reply Collective. Lund University.
- Future of Life Institute: <https://futureoflife.org/>
- Gans, J. (2018). *Self-Regulating Artificial General Intelligence*, NBER Working Paper No. 24352, February 2018, JEL No. C72, D02
- Garreau, J. (2005). *Radical Evolution (The Promise and Peril of Enhancing Our Minds, Our Bodies - and What It Means to Be Human)*, Broadway Books.
- Gazzaniga, M. (1967). The split brain in man. *Scientific American*, 217(2), 24–29.
- Gendler, T. (1996). *Imaginary exceptions: On the powers and limits of thought experiments*.
- Gendler, T. (1998). *Continence on the cheap: a response to Roy Sorensen*. *Mind*. 107:821.
- Gendler, T. (1998). Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science*. 49:397-424.
- Gendler, T. (2000). *Thought Experiment: On the Powers and Limits of Imaginary Cases*, London: Routledge.

- Gendler, T. (2004). Thought experiments rethought—and reperceived, *Philosophy of Science* 71: 1152–1163.
- Goertzel, B. (2006). *The hidden pattern: A patternist philosophy of mind*, Brown Walker Press.
- Goertzel, B. (2007). *The Role of AGI in Cybernetic Immortality*, The Journal of Geoethical Nanotechnology.
- Goertzel, B. (2010). *AI for Increased Human Healthspan*. Next Big Future, August 14.
- Goertzel, B. (2012). *Why an Intelligence Explosion is Probable (Singularity Hypothesis)*, Springer Verlag, Berlin.
- Goertzel, B. (2015). Superintelligence: Fears, Promises and Potentials (Reflections on Bostrom’s Superintelligence, Yudkowsky’s From AI to Zombies, and Weaver and Veitas’s “Open-Ended Intelligence”), *Journal of Evolution and Technology* – Vol. 24 Issue 2 – pgs 55-87.
- Goertzel, B. (2016). Infusing Advanced AGIs with Human-Like Value Systems: Two Theses, *Journal of Evolution & Technology*, Vol. 26 Issue 1, pgs 50-72.
- Goertzel, B. and J. Pitt (2012). Nine Ways to Bias Open-Source AGI Towards Friendliness. *Journal of Evolution and Technology* 22(1): 116-131.
- Goertzel, B. (2015). Superintelligence: Fears, promises and potentials. *Journal of Evolutionary Technology*. 25, 55–87.
- Goertzel, B., and Pennachin, C. (2007). *Artificial General Intelligence*. Springer.
- Goldin, O. & Kilroe, P. (1997). *Human Life and the Natural World*. Readings in the History of Western Philosophy. Broadview Press.
- Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine, Franz L. Alt and Morris Rubinoﬀ, ed., *Advances in Computers (Academic Press)* 6: 31–88.
- Gooding, D. (1992). *What is Experimental about Thought Experiments?* in: Proceedings of the Biennial Meeting of the Philosophy of Science Association. Volume 2, Chicago: University of Chicago Press, 280-290.

- Gould, S.J. (1989). *Wonderful life: the burgess shale and the nature of history*. New York: Norton.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017) *When will AI exceed human performance? Evidence from AI experts*, arXiv:1705.08807.
- Greene et al, (2013). Metformin retards aging in *C. elegans* by altering microbial folate and methionine metabolism. *Cell* 153, 1, 228–239.
- Greene, J. (2016). *Solving the Trolley Problem, A Companion to Experimental Philosophy*, First Edition. Edited by Justin Sytsma and Wesley Buckwalter. © 2016 John Wiley & Sons, Ltd.
- Greenfield, Susan. (2012). The Singularity: Commentary on David Chalmers. *Journal of Consciousness Studies* 19 (1–2): 112–118.
- Groß, Dominik. (2009). Blessing or Curse? Neurocognitive Enhancement by Brain Engineering. *Medicine Studies* 1 (4): 379–391.
- Grunwald, A (2007). *Nanotechnology – Steps Towards Understanding Human Beings as Technology?*, NanoEthics, Springer Science.
- Gubrud, Mark Avrum. (1997). *Nanotechnology and International Security*. Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, November 5–8.
- Häggström, O. (2016). *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.
- Häggström, O. (2016). Aspects of mind uploading, Chalmers University of Technology and the Institute for Future Studies.
- Häggström, O. (2017). *The AI meeting in Brussels last week*. Häggström Hävdar
- Häggström, O. (2018). *A spectacularly uneven AI report*. Häggström Hävdar
- Häggvist, S. (2009). A model for thought experiments. *Canadian Journal of Philosophy* 39 (1): pp. 55-76
- Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon.

- Hanson, R. (2008). *Economics of brain emulations*. In Healey P. (Ed.) Tomorrow's people - proceedings of the james martin institute's first world forum: EarthScan.
- Hanson, R. (1994). *If uploads come first: The crack of a future dawn*. Extropy, 6.
- Hanson, R. (2004). *The next really big enormous thing*. Future Brief.
- Hanson, R. (2008). *Economics of the singularity*. IEEE Spectrum, pages 36–42,
- Hanson, R. (2016). *Age of Em: Work, Love and Life when Robots Rule the Earth*, Oxford University Press, Oxford.
- Harari, Y. (2016). *Homo Deus, A brief history of tomorrow*, Harvill Secker.
- Harris, J. (2012). *Fractal Architecture: Organic Design Philosophy in Theory and Practice*, Univ of New Mexico Pr.
- Haraway, D (1991). *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century in Simians, Cyborgs and Women: The Reinvention of Nature*, New York, Routledge.
- Harreveld et al (eds) (2016). *Constructing methodology for qualitative research. Researching education and social practices*. Palgrave studies in education research methods. Palgrave Macmillan, London.
- Hauser, L. (1997). Searle's Chinese Box: Debunking the Chinese Room Argument. *Minds and Machines*, Vol. 7 (1997), No. 2, pp. 199-226
- Hauskeller, M. (2012). My brain, my mind, and I: Some philosophical assumptions of mind uploading, *International Journal of Machine Consciousness*, 4 (1), pp. 187–200.
- Hauskeller, M. [2009] Making Sense of What We Are, *Philosophy* 84(1), 95-109.
- Hayles, N.K. [1996] How Cyberspace Signifies: Taking Immortality Literally, in G. Slusser et al. (eds.), Immortal Engines. Life Extension and Immortality in *Science Fiction and Fantasy* (University of Georgia Press, Athens and London), 111-121.
- Hayworth, K. (2010). *Killed by bad philosophy: Why brain preservation followed by mind uploading is a cure for death*. Essay published online at <http://brainpreservation.org/content/killed-bad-philosophy>

- Hayworth, K. (2012). Electron imaging technology for whole brain neural circuit mapping. *International Journal of Machine Consciousness*, 4(1).
- Hayworth, K.J. (2011). *Lossless thick sectioning of plastic-embedded brain tissue to enable parallelizing of SBFSEM and FIBSEM imaging*, High Resolution Circuit Reconstruction conference 2011, Janelia Farms, Ashburn, VA.
- Hayworth, K.J., Kasthuri, N., Hartweg, E. and Lichtman, J.W. (2007). *Automating the collection of ultrathin brain sections for electron microscopic volume imaging*. Program No. 534.6, 2007 Neuroscience Meeting, San Diego, CA.
- Heersmink, R. (2013). A taxonomy of cognitive artifacts: Function, information, and categories, *Review of Philosophical Psychology*, 4 (3), pp. 465–481.
- Heidegger, M. (1956 [1947]). *The Way back into the Ground of Metaphysics*, in Walter Kaufmann (ed.), *Existentialism from Dostoevsky to Sartre* (New York: World Publishing Co.).
- Heidegger, M. (1962 [1927]). *Being and Time*. Trans by John Macquarie and Edward Robinson (New York: Harper and Row Publishers).
- Hilbert, M. & López, P. (2011). *The World's Technological Capacity to Store, Communicate, and Compute Information*, Science.
- Hillerbrand, R. (2008). *Scale separation as a condition for quantitative modelling. Why mathematics works for some problems and fails for others*. Synthese.
- Hinchey et al, (2008). Software engineering and Formal Methods, *Commun, ACM* 51 (9): 54-59.
- Hochberg, L. (2006). *Neuronal ensemble control of prosthetic devices by a human with tetraplegia*, *Nature* 442, 164-171.
- Hofstadter, D. and Dennett, D. (1981). *The Mind's I*, Basic Books, New York.
- Hongladarom, S. (2011). Personal identity and the self in the online and offline world, *Minds and Machines*, 21, pp. 533–548.
- Hopkins, P (2005). *Transcending the Animal: How Transhumanism and Religion Are and Are Not Alike*, *Journal of Evolution and Technology*.

- Howard, A. (2000). *Philosophy for Counselling and Psychotherapy: Pythagoras to Postmodernism*. Macmillan International Higher Education.
- Hunt, T. et al (2019). *The easy part of the Hard Problem: A resonance theory of consciousness*, *Frontiers in human neuroscience*.
- Hutter, M. (2010). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer. Berlin.
- Hutter, M. (2012). Can Intelligence Explode? *Journal of Consciousness Studies*, 19:1-2.
- I. J. Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6:31–88, 1965.
- Ihde, D. (2011). *Of which Human are we post?* Metanexus.
- Iwuagwu, E. (2017). Martin Heidegger and the question of being, *Journal of Integrative Humanism*, Vol.8 No.1, p25-48.
- Jacobstein, N (2017). *On The Future of AI*, on Ted-Talk.
- Johnston, I. (2017, April 17). *Scientists invent mind-reading machine that turns your thoughts into words*. The Independent. Retrieved from <http://www.independent.co.uk/news/science/read-your-mind-brain-waves-thoughts-locked-in-syndrome-toyohashi-japan-a7687471.html>
- Josh, (2009). Uploading your mind, Human Enhancement and Biopolitics. *Journal of Machine Consciousness*, 4 (1), pp. 141–158.
- Joy, B. (2000). *Why the future doesn't need us*. Wired, 8.04. URL: http://www.wired.com/wired/archive/8.04/joy_pr.html.
- Kandel, Eric R., James H. Schwartz, and Thomas M. Jessell, eds. (2000). *Principles of Neural Science*. New York: McGraw-Hill.
- Karpathy, A. (2015). *The unreasonable effectiveness of recurrent neural networks*. Retrieved from <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>
- Ketele, J. M. & Roegier, X. (1999). *Metodologia de recolha de dados. Fundamentos dos métodos de observações, de questionários, de entrevistas e de estudo de documentos*. Instituto Piaget, Lisboa.

- Kim, J., Biessmann, F., & Lee, S. (2014). Reconstruction of hand movements from EEG signals based on non-linear regression. In *Brain-Computer Interface (BCI), 2014 International Winter Workshop on* (pp. 1–3). IEEE.
- Kirsh, D (2006). Distributed cognition: A methodological note, *Pragmatics & Cognition*, 14 (2), pp. 249–262.
- Klein, B. (2003), *Building a Bridge to the Brain*, available online:
<http://www.imminst.org/forum/index.php?act=ST&f=67&t=938&s>.
- Knott, G., Marchman, H., Wall, D., & Lich, B. (2008). Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *The Journal of Neuroscience*, 28(12), 2959–2964.
- Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach* (Roberts and Co.).
- Koch, C. (2019). *The Feeling of Life Itself, Why Consciousness Is Widespread but Can't Be Computed*, MIT Press.
- Koch, T (2010). *Enhancing Who? Enhancing What?*, Ethics, Journal of Medicine and Philosophy. Toronto.
- Koene, R. (2012, β1). Experimental Research in Whole Brain Emulation: The Need for Innovative In-Vivo Measurement Techniques. *Special Issue of the International Journal of Machine Consciousness* 4(1). (also available at carboncopies official site: <http://ccsandbox.carboncopies.org/experimental-research-brain-emulation/>)
- Koene, R. (2012). *Embracing competitive balance: The case for substrate-independent minds and whole brain emulation, Singularity Hypotheses - A Scientific and Philosophical Assessment*, Springer Verlag.
- Koene, R. (2014 β2). *Interview by Luke Muehlhauser to Randal Koene on whole brain emulation*, March 20, 2014.
<https://intelligence.org/2014/03/20/randal-a-koene-on-whole-brain-emulation/>
- Koene, R. (2012b). *Toward tractable AGI: Challenges for System Identification in Neural Circuitry* In Bach, J.; Goertzel, B.; and Ikle', M., eds., *Artificial General Intelligence*. 5th International Conference, AGI 2012, 136–147.

- Koene, R., & Deca, D. (Apr 2014). Whole Brain Emulation seeks to implement a mind and its general intelligence through system identification, *Journal of Artificial General Intelligence*, 4(3), pp. 1–9
- Koene, R.A. (2011a). *Pattern survival versus Gene survival*, KurzweilAI.net, February 11, 2011. <http://www.kurzweilai.net/pattern-survival-versus-gene-survival>
- Koene, R.A. 2013. *Uploading to substrate-independent minds*. In The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future, ed. M. More and N. Vita-More, 146–56. Malden, MA: Wiley-Blackwell.
- Kolb, B (2011). *Brain Plasticity and Behaviour in the Developing Brain*, Journal of The Canadian Academy of Child and Adolescent Psychiatry, Department of Neuroscience,, University of Lethbridge, Lethbridge, Alberta
- Kompridis, N. (2009) Technology's challenge to democracy: What of the human?, *Parrhesia*, 8, pp. 20–33.
- Koza, J. (2010). Human-Competitive Results Produced by Genetic Programming. *Genetic Programming and Evolvable Machines* 11 (3–4): 251–284.
- Kryder, M. H., and Chang Soo Kim. (2009). *After Hard Drives—What Comes Next?*, IEEE Transactions on Magnetics 2009 (10): 3406–3413.
- Kuhn, T. S. (1977). A Function for Thought Experiments. In T. S. Kuhn (Editor), *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: UniversityofChicagoPress,pp.240-265.
- Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. New York, NY: Penguin Books.
- Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. New York, NY: Penguin Press.
- Kurzweil, R. (2012). *How to Create a Mind, the secret of human thought revealed*, Viking Penguin.
- Laakasuo, M. (2018). *What makes people approve or condemn mind upload technology?*, PALGRAVE COMMUNICATIONS | (2018)4:84 | DOI: 10.1057/s41599-018-0124-6 | www.nature.com/palcomms 1

- Legg, S. & Hutter, M. (2007). *Universal Intelligence- A definition of Machine Intelligence*.
- Lem, S. (1957). *Dialogi*, Wydawnictwo Literackie, Krakow. English translation by Frank Prengel at LEM.PL, <http://english.lem.pl/index.php/works/essays/dialogs/106-a-look-inside-dialogs>.
- Lenhard, J. (2018). *Thought Experiments and Simulation Experiments: Exploring Hypothetical Worlds*, Bielefeld University. In: Brown, J.R.; Fehige, Y. and Stuart, M.T., eds.: The Routledge Companion to Thought Experiments. Routledge, London & New York, 2018: 484-497
- Levin, J. (2013). *Functionalism*. In E. N. Zalta (Ed.) The Stanford encyclopedia of philosophy (fall 2013 edition).
<http://plato.stanford.edu/archives/fall2013/entries/functionalism/>
- Lewis, C. (1956). *Mind and the World-order: Outline of a Theory of Knowledge*, Courier Corporation.
- Li, Yu, & Liao (2001). Chaos in a three-neuron hysteresis hopfield-type neural networks. *Physics Letters A*, 285: 368-372, 2001.
- Lieber, C (2016). *Nano-Bioelectronics*, Chemical Reviews.
- Linssen, C. et al (2016). *Embodiment in Whole-Brain Emulation and its Implications for Death Anxiety*. Journal of Evolution & technology.
- Ljung, L. (2008). *Perspectives on system identification*. In In Plenary talk at the proceedings of the 17th IFAC World Congress, Seoul, South Korea.
- Loosemore, R. (2014). *Qualia Surfing*. In Intelligence Unbound: The Future of Uploaded and Machine Minds, John Wiley & Sons, p231-239
- Lytton, William W. (2002). *From Computer to Brain: Foundations of Computational Neuroscience*. New York: Springer.
- Markoff, J. (2013). *Obama Seeking to Boost Study of Human Brain*. The New York Times: 17 Feb, 2013.
- Markram, H. (2011). *Introducing the Human Brain Project*, Elsevier.
- Markram, H. (2013). Seven challenges for neuroscience, *Functional Neurology*, 28(3): 145 – 151.

- Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, vol.7:153-160.
- Martins, N.R.B., Erlhagen, W. & Freitas, R.A. (2012). Non-destructive whole-brain monitoring using nanorobots: Neural electrical data rate requirements, *International Journal of Machine Consciousness*, 4 (1), pp. 109–140.
- McCormick, B. and Mayerich, D.M. (2004). *Three-Dimensional Imaging Using Knife-Edge Scanning Microscopy*. In Proceedings of the Microscopy and Micro-analysis Conference 2004, Savannah, GA.
- McGin, Edelman and Tononi, (2000). *A Universe of Consciousness: How Matter Becomes Imagination*, Basic Books.
- McKibben, B (2003). *Enough - Staying Human In An Engineered Age*, Henry Holt & Company, New York.
- McIntyre, R. (2015). Aldehyde-stabilized cryopreservation, Elsevier, *Cryobiology*, Volume 71, Issue 3, pp 448-458, Medical Productions Inc. USA.
- Menary, R. (2007) *Cognitive Integration: Mind and Cognition Unbounded*, London: Palgrave Macmillan.
- Merkle, R. (1994). *The molecular repair of the brain*. Cryonics magazine, 15.
- Merkle, R. (1989). *Large scale analysis of neural structures*. Palo Alto, California: Xerox Palo Alto Research Center. Coyote Hill Road Palo Alto, CA 94304, 1989.
- Merkle, R. (1992). The technical feasibility of cryonics, *Medical Hypotheses* 39, 6-16.
- Milano, G. et al (2018). *Recent Developments and Perspectives for Memristive Devices Based on Metal Oxide Nanowires*, Advanced Electronic Materials.
- Milano, G. et al (2018). Self-limited single nanowire systems combining all-in-one memristive and neuromorphic functionalities, *Nature Communications* 9, Article number: 5151 (2018)
- Minsky, M. (1986). *The Society of Mind*, Simon and Schuster, New York.
- Mitchison, A. (2011). *Kim Noble: The woman with 100 personalities*, Guardian, 30 Sept 2011, [Online], <http://www.guardian.co.uk/lifeandstyle/2011/sep/30/kim-noble-woman-with-100-personalities>.

- Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, Cambridge, MA.
- Moravec, H. (1976). *The Role of Raw Rower in Intelligence*. Unpublished manuscript.
- Morris, S. (2013). *Mind uploading and identity*, Singularity Weblog, Danaylov, N. (ed.).
<https://www.singularityweblog.com/mind-uploading-and-identity>
- Moskowitz, C. (2016). *Are we living in a computer simulation?* Scientific American. <https://www.scientificamerican.com/article/are-we-living-in-a-computer-simulation/>
- Muehlhauser & Salamon, (2012). *Intelligence explosion: evidence and import*. In: Eden Amnon, Søraker Johnny, Moor James H, Steinhart Eric, editors. *Singularity hypotheses: a scientific and philosophical assessment*. Springer; 2012. p. 15e42.
- Müller, V. (2016). *Risks of artificial intelligence*, (London: CRC Press - Chapman & Hall)
- Mulgan, T. (2015). *Purpose in the Universe: The moral and metaphysical case for Ananthropocentric Purposivism*, Oxford University Press.
- Nagel, T. (1971). Brain bisection and the unity of consciousness. *Synthese*, 22, 396–413.
- Nagy, Béla, J. Doyne Farmer, Jessika E. Trancik, and Quan Minh Bui. 2010. *Testing Laws of Technological Progress*. Santa Fe, NM: Santa Fe Institute, September 2
- Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, Mass: MIT Press.
- Nersessian, N. (2018). *Cognitive Science, Mental Modeling, and Thought Experiments*. In: Brown, J.R.; Fehige, Y. and Stuart, M.T., eds.: *The Routledge Companion to Thought Experiments*. Routledge, London & New York, 2018.
- Norman, D. (1991). *Cognitive artifacts*, in Carroll, J.M. (ed.) *Designing Interaction: Psychology at the Human–Computer Interface*, Cambridge: Cambridge University Press.
- Norton, J. (1996). Are Thought Experiments Just What You Thought? *Canadian Journal of Philosophy*. Vol 26, Issue 3, pp. 333-366
- Obama, B. (2013). *the “BRAIN” Initiative*, <https://obamawhitehouse.archives.gov/BRAIN>

- Oizumi M, Albantakis L, Tononi G, (2014). *From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0*. PLoS Comput Biol 10(5): e1003588. doi:10.1371/journal.pcbi.1003588, Editor: Olaf Sporns, Indiana University, United States of America
- Olson, E. (2010). *Personal identity*, The Stanford Encyclopedia of Philosophy, Zalta, E. N. (ed.).
- Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence* 26(3): 303-315.
- Omohundro, S. 2008. *The basic AI drives*. In Proceedings of the First AGI Conference, Volume 171, Frontiers in Artificial Intelligence and Applications, edited by P. Wang, B. Goertzel, and S. Franklin, IOS Press.
- Oncel, D., Demetriades, D., Gruen, P., et al. (2007). Brain lobectomy for severe head injuries is not a hopeless procedure, *Journal of Trauma*, 63(5), 1010–1013.
- Parfit, D. (1984). *Reasons and Persons*, Oxford University Press, Oxford.
- Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press.
- Pigliucci, M. (2014). *Uploading: a philosophical counter-analysis*, in Blackford and Broderick, pp 119-130.
- Pihlaja, R. (2008). *Joined for Life: Abby and Brittany Turn 16*, (Film), Advanced Medical Productions Inc. USA.
- Pohl, J. (2015). *Artificial Superintelligence: Extinction or Nirvana?* InterSymp-2015, 3 August, 2015 RESU107-IS15
- Posner, R. A. (2004). *Catastrophe: Risk and Response*. New York, NY: Oxford University Press.
- Preston, C (2005). *The Promise and Threat of Nanotechnology: Can Environmental Ethics Guide Us?*, ScholarWorks at University of Montana, Philosophy Faculty Publications.
- Pulsifer, M., Brandt, J., Salorio, C., et al. (2004). The cognitive outcome of hemispherectomy in 71 children. *Epilepsia*, 4(3), 243–254.

- Putnam, H. (1967). *Psychological predicates*. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
- Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.
- Richards et al. (2004). *Chips, Architectures and Algorithms: Reflections on the Exponential Growth of Digital Signal Processing Capability*. Unpublished manuscript.
- Roco and Bainbridge (2006). *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society*, Springer Verlag.
- Rojcewicz, Richard. *The Gods and Technology: A Reading of Heidegger*. Albany: State University of New York, 2006.
- Rothblatt, M. (2007). *Cybervest Project*, [Online], <http://www.cybervest.org>.
- Rothblatt, M. (2012). The Terasem mind uploading experiment. *International Journal of Machine Consciousness*, 4(01), 141–158.
- Rothblatt, M. (2013). *Techno-immortalism 101: What are mindfiles?* In Longevitize! *Essays on the science, philosophy & politics of longevity*, ed. F. Cortese, 62–66.
- Rothblatt, M. 2014. *Virtually human: The promise – and the peril – of digital immortality*. New York: St. Martin's Press.
- Rottman, J. et al, (2014). Evolution, Development, and the Emergence of Disgust, *Evolutionary Psychology*, <https://doi.org/10.1177/147470491401200209>
- Roux, S. (2011). *The Emergence of the Notion of Thought Experiments*. Brill, pp.1-36.
- Sagan, C. (1977). *The dragons of Eden*. New York: Random House.
- Salamon, A. (2012). *Intelligence Explosion: Evidence and Import*, Luke Muehlhauser, Machine Intelligence Research Institute.
- Sandberg, A. (2013). Feasibility of whole brain emulation. In Müller, V. C. (Ed.), *Theory and philosophy of artificial intelligence (SAPER)* pp. 251–64). Berlin: Springer.
- Sandberg, A. (2014). Ethics of brain emulations, *Journal of Experimental & Theoretical Artificial Intelligence*, Vol 26, Issue 3: Risks of General Artificial Intelligence.

- Sandberg, A. & Bostrom, N. (2008). *Whole Brain Emulation: A Roadmap, Technical Report #2008-3*, Future of Humanity Institute, Oxford University.
- Sandberg, K., et al (2010). *Measuring consciousness: Is one measure better than the other?*, Elsevier, *Consciousness and Cognition* 19, 1069–1078.
- Sartre, J-P. (1943). *L'Être et le néant*. Éditions Gallimard
- Sartre, J-P. (1957). *Existentialism and human emotions*. Citadel Press, Kensington Pub. Corp.
- Sartre, J-P. (1956 [1946]). *Existentialism is a Humanism*, in Walter Kaufmann (ed.), *Existentialism from Dostoevsky to Sartre*. (New York: World Publishing).
- Schaul, Tom, and Jürgen Schmidhuber. 2010. "Metalearning." *Scholarpedia* 5 (6): 4650.
- Schierwagen, A. (2011). *Reverse Engineering for Biologically Inspired Cognitive Architectures: A Critical Analysis*. In *From Brains to Systems: Brain-Inspired Cognitive Systems* edited by Carlos Hernández, Ricardo Sanz, Jaime Gómez-Ramirez, Leslie S. Smith, Amir Hussain, Antonio Chella, and Igor Aleksander, 111–121. *Advances in Experimental Medicine and Biology* 718. New York: Springer.
- Schmidhuber, Jürgen. 2007. "Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers." In Goertzel and Pennachin 2007, 199–226.
- Schmitt, C (2015). *Nanomaterials—Tools, Technology and Methodology of Nanotechnology Based Biomedical Systems for Diagnostics and Therapy*, Biomedicines.
- Schneider, S. (Mar 2014). The philosophy of 'Her', The New York Times Opinionator. *Science* 282, 1846–1851.
- Schrödinger, E. (1958). *What is Life? The physical aspect of the living cell & Mind and Matter*, Cambridge University Press, Cambridge.
- Searle, J. (1980). *Minds, Brains, and Computers* in *Minds, Brains and Programs* in *The Behavioral and Brain Sciences*, vol. 3. Copyright @ 1980 Cambridge University Press.
- Searle, J. (2004). *Mind: A Brief Introduction*, Oxford University Press, Oxford.

Sejnowski, J., et al (2019). Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning, Salk Institute for Biological Studies, La Jolla, CA, and approved March 19, 2019 (received for review July 28, 2018), PNAS (Proceedings of the National Academy of Sciences of the United States of America), available online:

<https://www.pnas.org/content/116/17/8554.abstract>

Sen, S. (2016). *Measuring Consciousness*, Journal of Consciousness Studies.

Seth, A.K. (2007). *Post-decision wagering measures metacognitive content, not sensory consciousness*. Conscious Cogn. DOI: 10.1016/j.concog.2007.05.008 (www.sciencedirect.com).

Shannon, C. (1948). A Mathematical Theory of Communication, *The Bell System Technical Journal*, 27, 379-423, 623-656.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2017). *Deep image reconstruction from human brain activity*. BioRxiv, 240317.

Shoemaker, S. (1984). *Personal identity: A materialist's account*. Shoemaker and swinburne, personal identity (pp. 69–132). Oxford: Blackwell.

Shulman, C., Sandberg, A. (2010). *Implications of a software-limited singularity*. Proceedings of the European Conference of Computing and Philosophy.

Sim Bamford and John Danaher (2017). Transfer of Personality to a Synthetic Human 'Mind Uploading' and the Social Construction of Identity, *Journal of Consciousness Studies*, 24, No. 11–12, pp. ?–?

Soltanian-Zadeh, S. (2019). *Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning*, Proc Natl Acad Sci USA. 2019 Apr 23;116(17):8554-8563. doi: 10.1073/pnas.1812995116. Epub 2019 Apr 11.

Sorenson, R. (1992). *Thought Experiments*, Oxford University Press.

Sowka, D. (2012). *Deus Ex Machina, Artificial General Intelligence and Responsibility*, Matrikel-Nr.: 1047850,

https://www.academia.edu/34926666/Responsibility_and_Artificial_General_Intelligence

- Stanley, R. (1999). Qualia space. *Journal of Consciousness Studies*, 6(1), 49–60.
- Strout, J. (Aug 1997). *The philosophy and technology of mind uploading*, Sci-Fi Arizona, Writer's Workshop. <http://www.scifi-az.com/articles/guest1.pdf>
- Swinburne, R. (1984). Personal identity: A dualist theory. *Shoemaker and swinburne, personal identity*, (pp. 1–66). Oxford: Blackwell.
- Takinoue, M. et al (2019). *DNA Origami Nanoplate-Based Emulsion with Nanopore Function*. Angewandte Chemie International Edition, 2019
- Tegmark, M. (2011). *The Future of Life: A Cosmic Perspective*, presented at the Singularity Summit 2011, Oct. 15, New York, NY.
- Tegmark, M. (2017). *Life 3.0 : being human in the age of artificial intelligence*, Alfred A. Knopf, New York, 2017.
- Thomas, J. (2016). In defense of philosophy: a review of Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, *Journal of Experimental & Theoretical Artificial Intelligence*, 28:6, 1089-1094,
- Tononi G, Koch C (2008). *The neural correlates of consciousness: an update*. Ann N Y Acad Sci 1124: 239–61.
- Tononi G, Laureys S (2009). *The neurology of consciousness: an overview*. The neurology of consciousness, 375–412.
- Tononi, G. (2004). *An information integration theory of consciousness*. BMC Neurosci. 5, 42
- Tononi, G. and Edelman, G.M. (1998). *Consciousness and complexity*. Science 282, 1846–1851
- Trafton, A. (2012). *Robots that reveal the Inner Workings of Brain Cells*. MIT News. Available electronically at <http://web.mit.edu/newsoffice/2012/robots-recording-neurons-0507.html>
- Trappenberg, T. (2009). *Fundamentals of Computational Neuroscience*. 2nd ed. New York: Oxford University Press.

- Tripp, S. and Grueber, M. (2011). *Economic Impact of the Human Genome Project*. Battelle Memorial Institute.
- Turchin, A. et al, (2017). Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence, *Informatica*41(2017)401–417.
- Turchin, A. (2018). *Digital Immortality: Theory and Protocol for Indirect Mind Uploading*, (c) Digital Immortality Now, DIN, 1987–2018
- Urban T. (2015a). *The AI Revolution: Our Immortality or Extinction*, Wait But Why Website (Tim Urban and Andrew Finn), January [http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html]
- Urban T. (2015b). *The AI Revolution: The Road to Superintelligence*, Wait But Why Website (Tim Urban and Andrew Finn).
[http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html]
- Van der Velde, F. (2010). *Where Artificial Intelligence and Neuroscience Meet: The Search for Grounded Architectures of Cognition*. Advances in Artificial Intelligence, no. 5.
- Venkatesen, T. et al (2020). Charge disproportionate molecular redox for discrete memristive and memcapacitive switching, *Nature Nanotechnology* volume 15, pages 380–389.
- Vinge, V. (1981). *True names and other dangers*, Baen Books, ISBN-13: 978-0671653637.
- Vinge, V. (1993), *The Coming Technological Singularity: How to survive in the post human era*. Vision-21: Interdisciplinary Science & Engineering in the Era of CyberSpace, proceedings of a Symposium held at NASA Lewis Research Center (NASA Conference Publication CP-10129).
- Vinge, V. (1983). *First word*. Omni, January 1983, p. 10.
- Vinge, V. (2008). *Signs of the Singularity*, IEEE Spektrum's Special Report.
- Wagner, J. (2017). *Is artificial intelligence dangerous? Self-improving systems and the intelligence explosion*, Term Paper, University of Heidelberg

- Waldner, JB. (2008). *Nanocomputers and Swarm Intelligence*, First published in Great Britain and the United States in 2008 by ISTE Ltd and John Wiley & Sons, Inc.
- Walter, C. (2005). *Kryder's Law*. Scientific American, July 25.
- Walters & Palmer (1997). *The Ethics of Human Gene Therapy*, Oxford University Press, New York.
- Warwick, K (2016). *Transhumanism: Some Practical Possibilities*. FifF Kommunikation, 2/2016, pp.24-27
- Weitz, K. (2006) Kim Noble: *A woman divided*, Independent, 27 Aug 2006, [Online], <http://www.independent.co.uk/news/people/profiles/kim-noble-a-woman-divided-413223.html>.
- Wiley & Koene (2015). *The Fallacy of Favoring Gradual Replacement Mind Uploading Over Scan-and-Copy*, available online:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2596460
- Wiley, K. (2018). *Mind Uploading and the Question of Life, the Universe, and Everything*, Carboncopies (<http://carboncopies.org>), This 2018 paper is the original version of a shorter article published by IEET in 2015 (Wiley 2015).
- Wiley, K. (2011a). *Implications of computerized intelligence on interstellar travel*. H+ Magazine. [http:// hplusmagazine.com/2011/09/23/implications-of-computerized-intelligence-on-interstellar-travel/](http://hplusmagazine.com/2011/09/23/implications-of-computerized-intelligence-on-interstellar-travel/)
- Wiley, K. (2011b). *The Fermi Paradox, self-replicating probes, and the interstellar transportation bandwidth*. arXiv:1111.6131.
- Wiley, K. (2014). *A taxonomy and metaphysics of mind-uploading*. Humanity+ Press and Alautun Press. Wiley, K. 2015. *Mind uploading and the question of life, the universe, and everything*. IEET.
<https://ieet.org/index.php/IEET2/more/wiley20150720>
- Wiley, K. (2014) *A Taxonomy and Metaphysics of Mind-Uploading*, Humanity+ Press and Alautun Press.
- Wiley, K. (2015). *Mind Uploading and The Question of Life, the Universe, and Everything*, Ethical Technology

- Wiltsche, H. (2016). *Phenomenology and Thought Experiments. Thought Experiments as Anticipation Pumps*, Department for Philosophy University of Graz, Austria
- Wing, J. (1990). A Specifier's Introduction to Formal Methods, *Computer* 23 (9): 8-23.
- Winner, L. (1977). *Autonomous Technology: Technics-out-if-Control as a Theme in Political Thought*. Cambridge, MA: The MIT Press, 1977.
- Yampolskiy, R. (2014). *The Universe of Minds*. arXiv preprint arXiv:1410.0369.
- Yampolskiy, R. and Fox, J. (2013). *Safety Engineering for Artificial General Intelligence*. *Topoi* 32(2): 217-226.
- Yampolskiy, R. (2012). *Leakproofing the Singularity: Artificial Intelligence Confinement Problem*. *Journal of Consciousness Studies* 2012 (1–2): 194–214
- Yudkowsky, E. (2001). Creating Friendly AI 1.0: The analysis and design of benevolent goal architectures. The Singularity Institute, San Francisco, CA, June 15.
- Yudkowsky, E. (2008). *Artificial intelligence as a positive and negative factor in global risk*. In *Global Catastrophic Risks*, pages 232–263. Oxford University Press.
- Yudkowsky, E. (2008). *Timeless identity*, Less Wrong, June 3.
- Yudkowsky, E. (2011). *Complex value systems in friendly AI*. In Schmidhuber, Thórisson, & Looks 2011, 388–393.
- Yudkowsky, E. (2012). *Friendly Artificial Intelligence*. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*. A. H. Eden, J. H. Moor, J. H. Soraker & E. Steinhart. New York, NY: Springer: 181-194.
- Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics*. Berkeley, California, Machine Intelligence Research Institute.
- Yudkowsky, E. (2017). *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.
- Zador, A. (2011). Sequencing the connectome: A fundamentally new way of determining the brain's wiring diagram, Project Proposal, Paul G. Allen Foundation Awards Grants.
- Zeigler, P. B., Praehofer, H., & Kim, T. (2000). *Theory of modeling and simulation: integrating discrete event and continuous complex dynamics systems*. Academic Press.

Zhang, F (2014). *Structural DNA nanotechnology: state of the art and future perspective*, Journal of the American Chemical Society.

Zhong et al. (2019). *In vivo molecular imaging for immunotherapy using ultra-bright near-infrared-*lib* rare-earth nanoparticles*, Nature Biotechnology, volume 37, pages1322–1331.

Webinars and Workshops

(available @ <https://carboncopies.org/past-events/>)

Webinar: Carboncopies Foundation 2018 Winter Workshop, Jan 28, 2018;
TRANSCENDING BIOLOGY: Reverse engineering the brain

Webinar: Carboncopies Foundation Summer 2018 Workshop, September 15, 2018;
METAPHYSICS AND ETHICS OF WHOLE BRAIN EMULATION

Webinar: “WBE - Roadmap 2018, Where are we now?”

Webinar: Carboncopies Foundation Winter 2019 Workshop, March 16, 2019; *AI SAFETY AND WHOLE BRAIN EMULATION.*

Webinar: Carboncopies Foundation Spring 2019 Workshop, June 1, 2019; *REVIEW OF THE OXFORD ROADMAP ON WHOLE BRAIN EMULATION.*

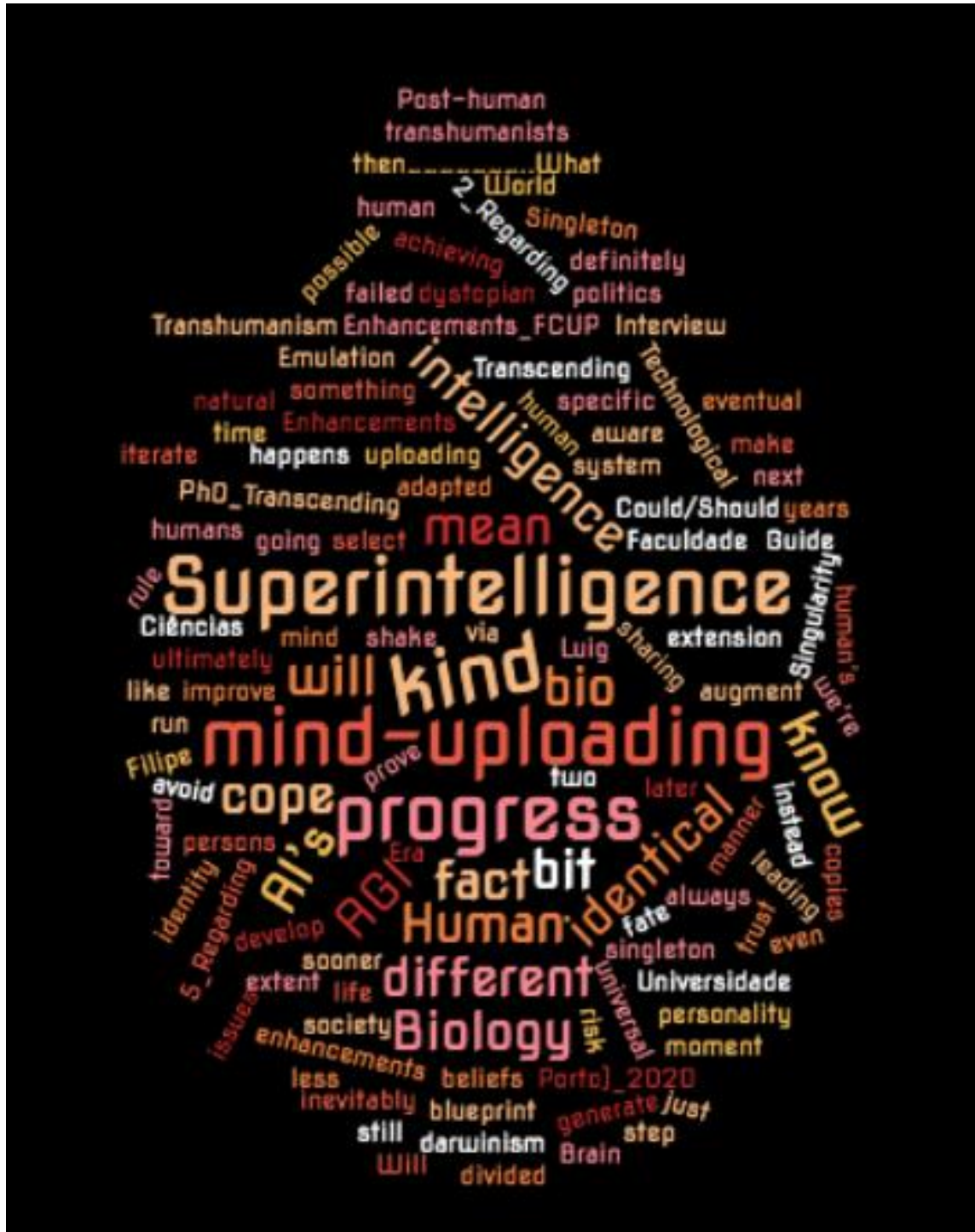
Webinar: Carboncopies Foundation Summer 2019 Workshop, September 21, 2019;
Updating the Roadmap to Whole Brain Emulation Part 2: *Where We Go From Here.*

Webinar: Carboncopies Foundation Fall 2019 Workshop, December 14, 2019; Updating the Roadmap to Whole Brain Emulation Part 3: *Consciousness.*

Webinar: Carboncopies Foundation Winter 2020 Workshop, March 15, 2020; Updating the Roadmap to Whole Brain Emulation Part 4: *SUCCESSFUL MIND UPLOADING. What is Personal Identity and How Can It Survive Mind Uploading*

APPENDIX

Semi-Structured Interviews



APPENDIX A _ Request for Interview

INTRODUCTION OF INTERVIEWER / CONSENT

Dear XX (*Stakeholders Name*),

My name is Filipe Luig, I am a PhD student at Faculdade de Ciências da Universidade do Porto, Portugal currently writing my thesis report on *Transcending Biology toward Human Enhancements*.

Selection for interview was based on your specific work in relation to Human Enhancements and its impact on what it means to be human. Interviews will be recorded to ensure accuracy, permission will be asked before using direct quotes from the interview. I hereby enclose a Data Confidentiality Agreement attached.

Opinions for response are flexible so please let me know which is best for you. It would represent a major contribution to my thesis if it would be possible to carry out the interview via Skype or by Phone. If you cannot find the time a written response is valued as well. I hope we can arrange an appointment at your convenience.

During the interview which would take around ½ Hour comprised of about 4-5 questions regarding the implications of human enhancements such as WBE, Intelligence Augmentation, its supersizing, or the creation/ development of a Superintelligence and the possibility of mind-uploading, but most specifically the impact on what it means to be human today and in a post-human Era. All additional general comments are welcome as well. Your contributions would be very inspiring. I would like to thank you in advance for your time and cooperation.

Best Regards,

Filipe Luig _ (Skype: nanosimbiotica / e-mail: nanosimbiotica@gmail.com)

APPENDIX B _ Candidates contacted for interview

Table 18 – List of Experts contacted for interview. (Luig, 2020)

ANDERS SANDBERG ____ ACCEPTED
ANNA SALAMON ____
AUBREY DE GREY ____
DAVID CHALMERS ____ ACCEPTED
DANIEL DENNETT ____ DECLINED

DIANA DECA ____
ERIC DREXLER ____ DECLINED
ELIZER YUDKOWSKY ____
FRANCIS FUKUYAMA ____
GERD LEONHARD ____ DECLINED

HENRY MARKRAM ____
JAMES BARRAT ____ ACCEPTED
JOEL GARREAU ____
KEITH WILEY ____
KEVIN WARWICK ____

LOUIS DEL MONTE ____
MARCUS HUTTER ____ DECLINED
MAX MORE ____
MAX TEGMARK ____
MICHAEL CERULLO ____

NATASHA VITA-MORE ____
OLLE HAGGSTROM ____ DECLINED
DENNIS BRAY ____
DAVID PEARCE ____ ACCEPTED
RANDAL KOENE ____ ACCEPTED

RAY KURZWEIL ____
VERNOR VINGE ____ DECLINED
YUVAL HARARI ____
NICK BOSTROM ____ DECLINED
ROBERT A. FREITAS ____

ROBIN HANSON ____ ACCEPTED
RICK CATTELL ____ ACCEPTED

TOM KOCH ____ ACCEPTED
NICHOLAS AGAR ____ ACCEPTED
LUKE MUELHAUSER ____

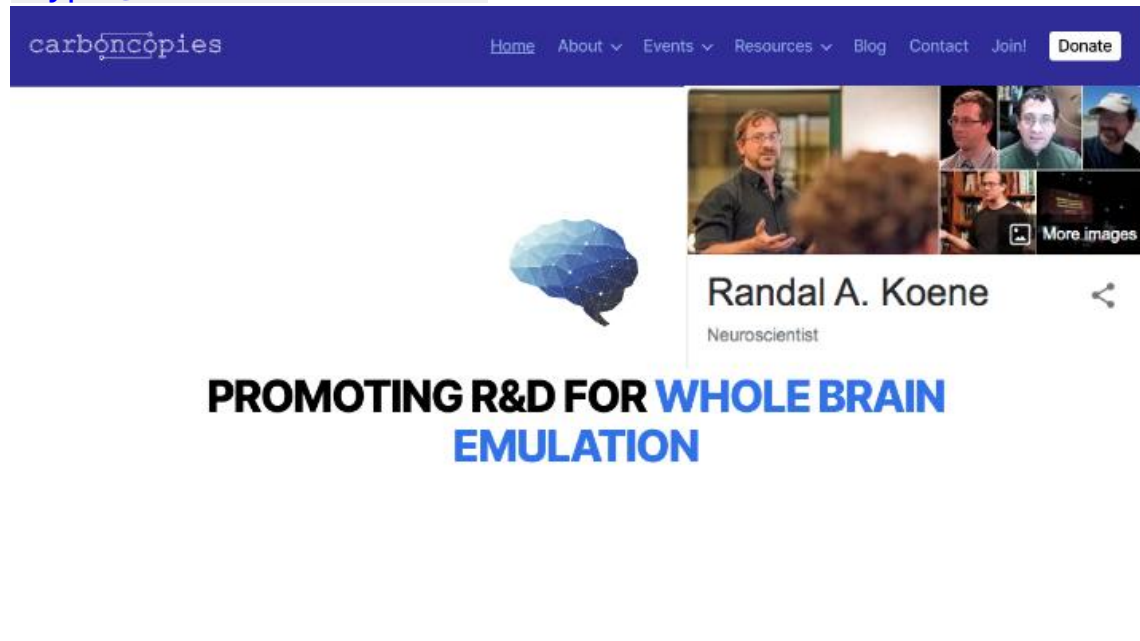
ROMAN YAMPOLSKIY ____
RICHARD LOOSEMORE ____ ACCEPTED

* Empty brackets correspond to Experts not having answered back my interview request.

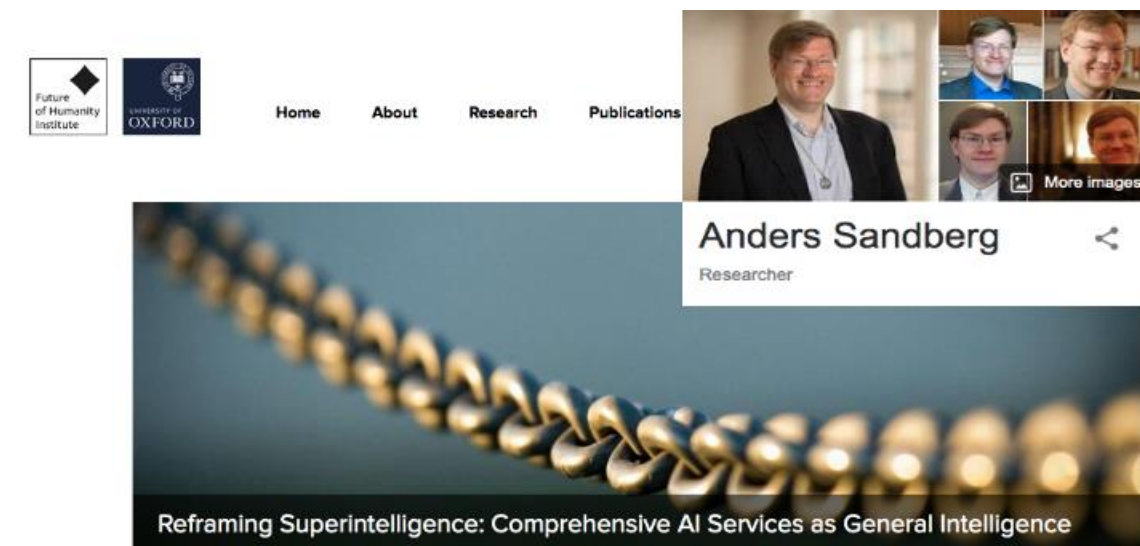
APPENDIX C _ Interviewees' Presentation

1 _ RANDAL KOENE _ CARBONCOPIES, USA

[skype@01MAR2020 ~50min](#)



2 _ ANDERS SANDBERG _ FUTURE OF HUMANITY INSTITUTE, UK [skype@11MAR2020 ~60min](#)



3 _ KENNETH HAYWORTH, BRAIN PRESERVATION FOUNDATION, USA

[skype@28MAR2020 ~60min](#)



Founded	27 August 2010 ^[1]
Founder	Kenneth Hayworth and John Smart ^[2]
Type	United States IRS exemption status: 501(c)(3), ruling year 2010 ^[1]
Focus	Brain preservation
Area served	Global
Key people	President Kenneth Hayworth ^[2]
Website	www.brainpreservation.org



Kurzweil AI
 The strange neuroscience of immortality ◀

4 _ DAVID CHALMERS, AUS [e-mail@?](#)

Books



The Conscious Mind
1996



The Character of Consciousness
2010



Constructing the World
2012



Philosophy of Mind: Classical and Contemporary Views
2002

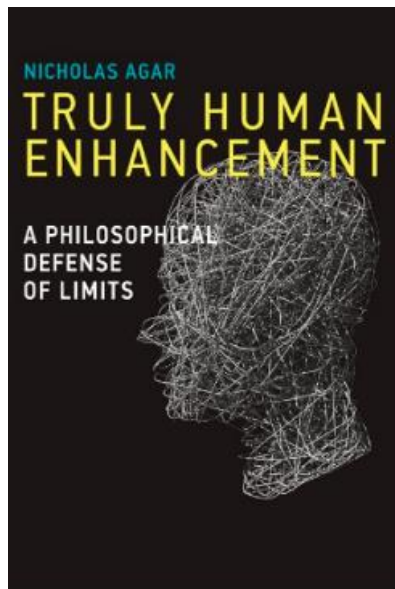


David Chalmers
 Australian philosopher

More images

5 _ NICK AGAR, VICTORIA UNIVERSITY, NZ

[skype@28MAR2020 ~60min](#)



From Basic Bioethics

Truly Human Enhancement A Philosophical Defense of Limits

By Nicholas Agar

A nuanced discussion of human enhancement that argues for enhancement that does not significantly exceed what is currently possible for human beings.

**Nicholas
 Agar**
 Professor



5 _ JAMES BARRATT, FILMMAKER, USA

[skype@30MAR2020 -60min](#)

OUR FINAL INVENTION

Artificial Intelligence helps choose what books you buy, what movies you see, and even who you date. It's in your smart phone, your car, and it has the run of your house. It makes most of the trades on Wall Street, and controls our transportation, energy, and water infrastructure. Artificial Intelligence is for the 21st century what electricity was for the 20th and steam power for the 19th.

But there's one critical difference — electricity and steam will never outthink you.

The Hollywood cliché that artificial intelligence will take over the world could soon become scientific reality as AI matches then surpasses human intelligence. Each year AI's cognitive speed and power doubles — ours does not. Corporations and government agencies are pouring billions into achieving AI's Holy Grail — human-level intelligence. Scientists argue



James Barratt
 Filmmaker



6 _ RICK CATTELL, SUN MICROSYSTEMS, USA

[skype @ 31MAR2020 ~30min + e-mail @38MAR](#)

Challenges for Brain Emulation: Why is Building a Brain so Difficult?

Rick Cattell, Alice C. Parker • Published 2012 • Engineering

In recent years, half a dozen major research groups have simulated or constructed sizeable networks of artificial neurons, with the ultimate goal to emulate the entire human brain. At this point, these projects are a long way from that goal: they typically simulate thousands of mammalian neurons, versus tens of billions in the human cortex, with less dense connectivity as well as less-complex neurons. While the outputs of the simulations demonstrate some features of biological neural networks... [CONTINUE READING](#)

[View PDF](#)

[Save To Library](#)

[Create Alert](#)

[Cite](#)

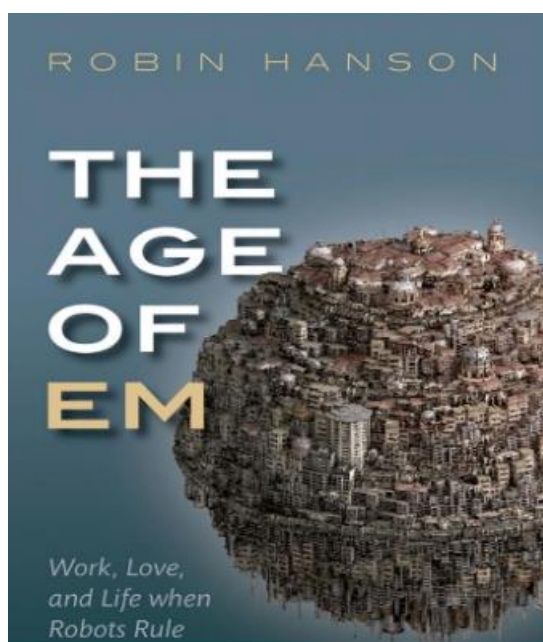
[Launch Research Feed](#)

Roderic
Geoffrey Galton
Cattell



7 _ ROBIN HANSON, FUTURE OF HUMANITY INSTITUTE, UK

[skype@01APR2020 ~60min](#)



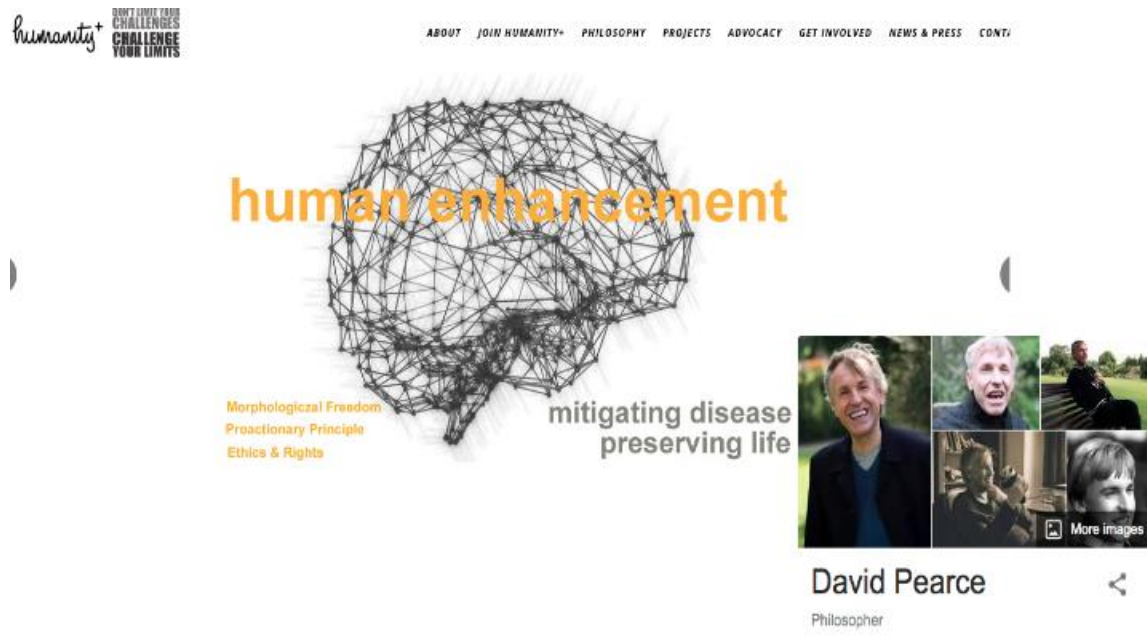
Robin Hanson

Professor



8 _ DAVID PEARCE, HUMANITY + , UK

[skype@03APR2020 ~60min](#)



10 _ TOM KOCH, BRITISH COLUMBIA UNIVERSITY, CAN

[skype@31MAR2020 ~60min](#)

Enhancing Who? Enhancing What? Ethics, Bioethics, and Transhumanism

Article (PDF Available) in [Journal of Medicine and Philosophy](#) 35(6):685-99 · November 2010 with 1,851 Reads ⓘ

DOI: 10.1093/jmp/jhq051 · Source: PubMed

[Cite this publication](#)



Tom Koch

159.27 · University of British Columbia - Vancouver

11 _ RICHARD LOOSEMORE, NY, USA

[skype @ 21APR2020 ~75min](#)



**Institute for Ethics and
Emerging Technologies**

BLOG

EVENTS

ABOUT

PROGRAMS

PUBLICATIONS

STAFF

Richard Loosemore



skype@01MAR2020~50min



FILIPPE _ Just to say a few words a quick intro about myself. I'm a biologist. For about 20 years now. I work in organ transplantation as a molecular biologist. After majoring in biology. I made a masters and two unfinished PhDs in nanotechnologies and nanosciences. Nevertheless, it gave me the opportunity to be part of the development of a couple of biochips, one of them a nanosensor device actually quite successful for detecting viruses. This could be a nice *timing* to talk about coronavirus, but that isn't the issue that brought us here together today. Anyway, this goes to say that I have this background in biotech and nanotech and linking these themes with the recent progress in AI is what basically permits the linkage of living stuff with non-living, bio and non-bio, which fascinates me, hence mind-uploading. So, there's the reason we are here today. Having that said, I've followed your work for quite some time now, I've read most of your papers, if not all, I've attended to the webinars of **carboncopies** too,

RANDAL _ Nice. Are you a member of the **carboncopies** facebook group as well?

FILIPPE _ No. Facebook?

RANDAL _ Oh, if it's not your 'thing' you know it's the oldest social page where we share things...

FILIPPE _ Yeah, as a fan of *carboncopies* I will definitely do that. OK, this being said I don't want to take too much of your precious time, I suggest we'd start with the interview, I've already sent you the interview guideline questions, if you can open here on skype, can you see it? You don't have to strictly follow the guide as it is, once again, just because of academic reasons, we just go from here, these 4-5 questions;

1_Regarding Whole Brain Emulation and mind-uploading

A simple one to start Q1: Would another me be me? (or just a copy of myself)

QUESTION I

RANDAL (laughs) _ A simple one?! This is not a simple one, this is the one where conversations usually get bogged down. We could side-step the issue for most people. What if we used a MU-procedure in which you don't copy yourself to some other thing

that is far away but instead you're replacing everything that is neurons and so you're putting artificial neurons in your head until you have an artificial brain, at no point was there ever a very clear distinction as to whether you switched over to a completely artificial brain so it feels like it still is me. Now, I say that's a way to deal with some of the questions there for most people because most of the time we don't think clearly about the issues, we don't think in great detail about it and so for example we don't notice that when we come up to that scenario that we're still talking about something where if you were to replace the neurons at a rate where you would actually get it done in your lifetime you would have to be replacing millions of neurons per second so even that gradual approach would still be something that would feel a lot like the whole copy scenario. This is something that Keith Wiley and I wrote about in a paper about the fallacy of trying to make this distinction between the copy scenario and the gradual replacement scenario...

FILIFE _ The gradual in-place replacement paper. I like that paper very much, yes...

RANDAL _ Yeah. Now the other part of this, the other side of it, is where you're getting bogged down in what are basically personal philosophies about consciousness and about identity. We tried to address that a little bit through recent workshops on consciousness and also the next one coming up about personal identity which goes specifically into the question of surviving uploading, so can you identity survival following MU. And the issue I think is partly that those of us who are functionalists, we might even say patternists, we tend to feel that first of all consciousness is something that is emergent from the brain, so that shouldn't be a problem. You can generate consciousness in the artificial brain just as much as you can in your own brain. And secondly, personal identity (this notion of who you are) isn't so much this in the moment-sense of '*oh I'm happy*', '*oh I'm sad*', that sort of thing, but it's how all of what we're experiencing right now gets processed through our memories through all of the synapses that have been changed over time since we were born, but even the settings of the synapses based upon whatever the blueprint was when we were born. The idea here being that let's say you, when you're happy, have a lot in common with when I'm happy. That sensation of being there in the moment and feeling happy that is something that humans share. Same thing is true for being sad, or for many other experiences, seeing the color blue, sure you can argue about whether you see the color blue the way I do. But there's a lot of these stuff that seem like a relatively standard circuit or process true across all humans and the thing that makes it unique that makes it me, that is my identity,

is the color that I give to it the experience that I go through because of all of the context that I have because of what it means to me. give to it because of the things that I've seen before that are blue and because of the experiences that come along with that. So, most of what it means to be you, your identity is encapsulated in your memory, not in your conscious momentary experience. So whether your experience right now turns off all of the sudden because you fell into icy water and your brain stopped working and then later someone brought you back to life...

FILIPPE _ Or in the case of a coma

RANDAL _ Yeah or because if it stops because you died and then you were stored in a brain preservation facility such as **Alcor** (More) or somewhere else, and then a hundred years later people decide to bring you back to life either biologically or to upload you, whatever these scenarios happen. What happens is that,

FILIPPE _ You maintain your pattern.

RANDAL _ Yeah, your pattern, that identity is still mostly there, if you can reconstruct those memories. And I say mostly there because I think the whole concept of identity being this sort of uniform this one thing that you can point to and say this is always me, is already a case of, i guess it's maybe not an illusion but it's kind of a wrong way to talk about it that we tend to be used to. So, for instance if i were to say I'm I still me now just like I was 5 minutes ago? Well I'm still mostly me from then. A few things might have changed but not much. But if I look back to my childhood a lot more has changed since then. Am I still me? Maybe a little bit, but I hardly remember anything from the time I was four (4). So,

FILIPPE _ And all of your atoms are different...

RANDAL _ Yeah. So, there is this aspect of identity that is this changing thing and now the thing that gets added to it, when you talk about something like MU is the possibility of having points in time where not only are you going from a previous to a next identity, you may also be branching! So it could be, if someone comes up with a MU technique that preserves the original brain in the process, we don't know of one of those yet, but we could end up with an artificial brain and a biological brain, and at the moment of creation - this is both you! But as soon as they start to experience slightly different things, they begin to become different, so you got a branching going on. Now, we don't really

know how you would do that yet because I don't know how you could upload without destroying the original brain in the process, so that doesn't happen. But that doesn't take away this question because you could still ask that once you made one upload but what if you suddenly made 10 copies? (the electronic version right), then you would have the same question: which of the ten is you? But the interesting thing I find is that when we talk about the upload some people have the same issue but very often when I talk to people they have a much easier time just being ok with the notion that there could be three (3) *yous* and they are all equally you. Because they're all digital versions anyway and they just became three digital versions. But then when it's about that initial copy, the copy from the biology to something digital, somehow there, it's a bigger problem at least it feels like a bigger problem, feels like something that how could this be true? And so I think there is a visceral reaction that sometimes goes back to a real personal philosophical difference and sometimes it's just visceral and just emotional. Just something we haven't really thought it through even if we actually do share the same philosophy. But I could say with certainty that even people who are long time avowed MU proponents and people working at **carboncopies** still struggle with these ideas from time to time. So that's not unusual at all.

FILIPPE _ OK, so you talked about the branching identity theory (another paper by you, yourself and Keith Wiley). So, beyond that and still in the realm of WBE and MU: Q2: What would represent the biggest risk in achieving something like this?

RANDAL _ What do you mean by achieving the biggest risk? To achieve it correctly or the biggest risk that exists once the technology is available?

FILIPPE _ Yeah, once the technology is available, what could it mean in terms of actually achieving those ten (10) digital copies of yours?

RANDAL _ Some of that risk depends a little bit on what a person or a society considers good versus bad, whatever we assign those criteria to. For example, I think many of us would find it bad (in other words it's a risk), if uploading were made possible and then what happens if you end up in a world where there are only 5 million copies of the same person and no others? Suddenly you have very little variation, differentiation, all of the other branches have basically ended and there's only just the one that proliferates. That would be sad, in a sense that it would be a sad outcome. Some people would say that there would be a risk involved if having MU available leads to either the eventual

disappearance of regular biological homo sapiens or if it leads to a situation where, because there are ten trillion uploads and only seven billion biological homo sapiens that in terms of the amount of influence that one or the other set of people have that it's huge to have both and the same could be said what if the mind uploads can all process things ten times faster and have other advances (e.g. mental advances), and develop mental possibilities that give them an edge? This could be perceived as a risk towards original biological humans. On the flip side though, the whole point, in my opinion, of going through MU is really the advancement of the species, in other words, I find it really hard for example to imagine that you would ever encounter a space-faring species that goes from star to star that is embodied as a sack of blood and tissue that needs to be oxygenated then something that is more suitable for the environment of space. So, I would imagine that all the really advanced species that do this have the technology like MU. I also think that the species that is advanced would want to be able to augment themselves and adapt themselves to different circumstances, because things don't always stay the same. At one point or another a giant asteroid is going to crash into the Earth. At some point or there's going to be a big gamma ray burst that does horrible things to the earth. There will be competition between artificial general intelligence and humans and the challenges will be different. It's not always going to stay the same. We need to be able to direct our own evolution into the directions where our chances are still being a part of what goes forward and taking advantage of possibilities and developments where our chances are simply better. And I think that's what really, when you look at what kind of a species in the universe would be really long-lived. One that would survive for, say, millions of years or billions of years and not just a few hundred thousand of years. Those would be the ones that have more adaptability and that have evolved themselves further. And this is a kind of super adaptability.

FILIPPE _ So, you say that we're kind of evolving into a more complex set of information. Like *Dataism*, as Yuval Harari has put so well in his book *Homo Deus*.

RANDAL _ Well I don't know if i would put those labels specifically. I only recently started reading that book actually. But I would say that I imagine we already have a situation right now where there isn't a single kind of intelligence on the Planet. There are people, there are governments and corporations, which are made of many people working together where you also have some kind of intelligence, we're beginning to see intelligent agents that come out of AI, that are not very that competitive yet, but they will become

more competitive. So, you end up having an ecosystem of intelligences and we fit in there somewhere. So, the question is what role we play in that ecosystem of intelligences, and that's why this kind of development is very important.

QUESTION II

FILIPE _ OK. This somehow can lead us to the next question regarding Superintelligence and the TS, if u will. If (when) it happens and we'd still be humans by that time, would we even understand it (be aware of it)?

RANDAL _ So many possible scenarios. I suspect that if you have a superintelligence that has its own agency that has its own goals, and that SuperIntelligence wants to create the sort of technology that will allow it to pursue those goals in the universe, that it will need to put together some kind of resources that we currently have available on earth. What sort of resources I don't know if it's just the coming technology or natural resources or energy I think that we would notice that I think that it is more likely that we would notice it than that we wouldn't notice it.

FILIPE _ Yeah, because as Yudkowsky said The AI doesn't love you nor does it hate you, but you are made of atoms it can use for something else...

RANDAL _ Yeah but also the AI is bound to the laws of physics and therefore every advance that it makes every improvement that makes in its process or whatever requires very sophisticated factory capabilities, and we would notice if things were being produced on this planet. So, Yes, we would be able to tell. Even if it's not taking a part of our atoms, that's kind of the extreme, even if it is simply tapping into our economy, and our infrastructure in order to be able to pursue its ends, we should be able to notice that.

FILIPE _ So can you envision that it would be possible for a singleton of some kind to rule the world?

RANDAL _ I guess theoretically yes I can see that. Theoretically there is at least one scenario where you could play out where you end up with a singleton controlling the world. Whether that singleton remains a singleton or whether you can then start speaking of the various competing algorithms inside its own brain as almost being the population of itself as its own species that becomes a matter of definition. You could even think of

yourself as collection of different brain processes and cognitive things that are going inside you. But, Yes, there being one agent from our perspective that ultimately controls our future (basically agent Smith of the matrix) I guess it's theoretically possible. I don't know if it's the most likely outcome, I actually don't think it's the most likely outcome because I think the world is very messy and there are so many actors in it. I'm not convinced that the outcome would be a singleton.

FILIPPE _ Would it be more probable by that time that we'd be already mixed, like a cyber sapiens of some kind?

RANDAL _ I don't know if we would be, I'm kind of hoping that we would be able to take advantage of AI ourselves and be sort of mixed in reducing the difference between us and our machines. But even if that's not the case, even if it's really AI separate from humanity, I think that there's a good chance that even the most advanced AI will find that it either has a competitor that is its own clone and therefore there are at least two or three of them perhaps pursuing slightly different aims, or something that was developed independently and that just happens to arrive at the same strength through a power struggle, a sort of arms race that you would get. I think that it's likely that there would be either multiple of that kind competing in a world or perhaps that they would actually drive each other to a halt and make it difficult for one to take off if they would do that. There are so many possible outcomes there. Singleton is a possibility, but it's only one branch with a certain probability attached to it.

QUESTION III

FILIPPE _ OK. Regarding Transhumanism and Posthumanism do you think that being human in a post-human era will be different than it is today?

RANDAL _ Being human?

FILIPPE _ Yes. In a Posthuman era we would probably be mixed with enhanced humans.

RANDAL _ So will being an original homo sapiens without being enhanced is that what you mean? So, coexisting with those who do have enhancements?

FILIPE _ Because you could envision the possibility that mind uploading would be technically feasible and then who gets to be uploaded and who doesn't? Who will govern this?

RANDAL _ So, the question as it is posed is will it be different? Wouldn't it be crazy to say no things would be the same, that would sound really weird of course I'm not going to say that It will be different. Of course, it's going to be. But, yes, you have other questions that are sort of subquestions right?

FILIPE _ Yes, in this case, if so, and we just covered that, then would this mean that we would be definitely "*more bit than bio*"? You know?

RANDAL _ Oh yes, ok! I think ultimately that seems very likely. And I think that because I, so that we can get past this point about bits versus analog: Some people claim that to create a proper processor like a brain you would need to build a neuromorphic chip that has basically an analog compute capability. But I think that that is a little bit silly, we have had these wars already, with compact discs versus analog music, you can always go to whatever degree of precision you'd want and it gets more reliable. So ultimately that's how they **..?** because they are more reliable than analog. And will we have more electronics or more software or more hardware that is developed and engineered versus evolved from biology? Yes, we will, because that can evolve faster and involve many more options. So, for example through the bits, you could choose any day the type of embodiment you would want to have. Different sensors, and different motor actuators whereas in your regular bio it might be more difficult to switch that up or to enhance it or choose what you want to do. So, I think ultimately Yes, I don't know at what pace but we'll be *more than bio*!

FILIPE _ OK. Could this also mean that an eventual cyber sapiens of some kind could be the next step in natural evolution? You align with that hypothesis?

RANDAL _ What is a cyber sapiens?

FILIPE _ A mixture of what we are today and that enhanced human we were talking about. It could be like having contact lenses today, but more enhancements, you see...

RANDAL _ So we might already be cyber sapiens in a sense...

FILIPE _ Well, at this dystopian kind of scenario these enhanced humans...

RANDAL _ So those who are completely uploaded or who have some attachments to machines that are giving them additional computational capabilities or sensor capabilities or something like that in a more direct way than what we have today. I think that is likely that that is something that is coming up. Whether I would call it the next step in human evolution? I think I wouldn't say that because that confuses us. Some people will say, well theoretically anything we do is a natural step because we are a product of nature therefore, everything we do is also natural. But that is just kind of redefining words because when we say natural evolution, what we typically mean is simply evolution changed through DNA because of environmental selection and selective pressure of natural selection. And, yes, we're departing from that somewhat. We still have selection pressures, but these aren't the same as in the natural environment and it's not through the change of DNA because of natural selection pressures and our normal procreation. So, it's a step of human's self-directed evolution, I would say. Not the next natural step in evolution, because that is just playing with words and giving them meanings, which is OK, you could do that, but it confuses people.

FILIPE _ Can you imagine the world of politics in ten or twenty years, divided between transhumanists against bioconservatives? So instead of having the right wing and the left wing?

RANDAL _ I wouldn't even say instead of. I think it is kind of there already. I think there is a possibility and it's actually something we should prepare for or be careful of, because it's already in the makings. I think the politics of today may confuse the issue a little bit because there are some groups that sort of belong to a party but lean over to the other. In the US on the democratic side, or the more left leaning side if you go to the extremes you'll find people who reject a lot of modern technology and they might even say oh no I do homeopathy I don't take medicine from pharmaceutical companies but on the right wing (if you look at the extreme side), you got the conservative religious groups, who also seem to not want to take that progress as much and are worried with the progress and with how the world has changing. Or perhaps people who are still engaged in activities where robots and machines are taking over their jobs. So you see meanwhile elsewhere in those groups you will see, I don't know - the libertarians who will sort of fit in with what is kind of a right wing right now but at the same time seem to align with the idea that technology will change things. So the current politics mixes it up a bit and makes it hard to see, but if you look at it carefully you can see the little things of that separation,

because ultimately when you look at what politics have been arguing about, it seems often about the pace of change. It seems that the more rapid pace of change causes a lot of concerns and groups of people anxious and it even threatens power structures that exist, and technological change tends to increase the rate of change or the pace of change in the world. So, I think a lot of the issues that play out in the political discourse are often ultimately based on exactly this sort of thing where this dividing shows up, and it's already hidden in the politics today, and it is probably going to emerge even more as progress...

FILIPPE _ And as these enhancements will be more ingrained in us, so people that have access to enhances such as MU and those who don't, this will split society and of course there will be parties that will defend those who don't get access to it. I would say that the richest will have more access to these enhancements.

RANDAL _ Yes, the distribution of availability of technology is definitely a problem we already see today because it doesn't have to be about human enhancements. It could be about something much simpler like access to medical technology and healthcare. So, the same problem exists already.

QUESTION IV

FILIPPE _ OK. So, the fourth and almost last question. And regarding SI and MU. I have a kind of thought experiment revolving MU: would it make sense to iterate intelligence via MU to augment or improve it and in this manner cope with AI's progress.

Let's take for instance the scenario where we could upload our minds and then expose our SIMs to various, millionfold different scenarios, economic, environmental, whichever, and just as we do today with DNA sequencing, we could artificially alter our SIMs, for instance SIM1 would be different copy, say just about 1% different from SIM99, and you would go on with these copies and find the one that would respond better to the different adversities or scenarios exposed, just like in a game. So, you could find the best chess-player, or go player, in our own minds...

RANDAL _ I think it's an interesting idea. I think it will apply to some problems as well as some skills better than others. I think that there are some problems as well as some brain regions that you could improve well through such a simulated evolutionary approach

rather than through direct engineering, because maybe you don't know the direction you are supposed engineer you don't have sufficient hints to tell you where you're supposed to go with it or you don't know how you would make the changes that don't make sense because you don't understand the system well enough. Those are all reasons where you would apply such an evolutionary approach such as that constant iteration where you would get some kind of random distribution of possibilities. If you have a better understanding of the system and how it works or if you have a better understanding of the problem and how to optimize towards a solution then often a more engineered approach is the faster one, even if subsets of that kind of engineering involve this kind of, what you might call deep-learning experimentation or a simulated annealing or whichever approach we would want use to apply a little bit of random to find the best path. So, I would say from an engineering perspective you would have to choose carefully when a certain method is going to be your most effective approach to an improvement in the output performance. So that is one part of my answer. The other part is that I would personally also want to make sure to apply these kinds of evolutionary approaches that are subjecting something to different environments very quickly and in many different settings. I would want to apply that only to unconscious pieces of a brain and never to an entire whole brain emulation, because if you believe that *you are real* when you are a copy, in other words, when you do not philosophically think that that is a nonconscious, nonthinking non-identity then subjecting all of these uploads to various scenarios is the same thing as subjecting people to them. That would be extremely unethical because in many cases that would be very uncomfortable to them. And even if you would volunteer you can't be sure that it's not an unethical situation once you get to part of that progress. So, I would want to be very careful with that. So, there's the ethical side and the other side is the engineering consideration. But yes, sometimes an evolutionary approach - good, sometimes - not optimal.

FILIPPE _ OK. Picking up exactly those words regarding identity and still in this kind of Thought Experiment, imagine the scenario we could make MU and then compare it (again making the analogy with DNA sequencing) you would obtain blueprints and I'm sure my DNA blueprint would be 99% identical to yours, this still doesn't mean we're the same person, of course not, you are Randal and I'm Filipe. But, nevertheless imagine this scenario where you could make the MU of different people and we would then just compare these mind blueprints and imagine the *mindblowing* case that we would find

out 100% identical blueprints of two different persons. Until what extent could we say that those are really two different persons if they share the identical MU blueprint?

RANDAL _ If their blueprints are completely identical? Then they would have to be the same person... The same as a copy of copy of copy... Sure!

FILIFE _ So you could via MU (if technically feasible) end up with identity issues and with individuality issues too...

RANDAL _ What do you mean by identity issues? Do you mean a Problem?

FILIFE _ No. Not a problem, I mean yes! Sure, it would stand as a problem if we would find out that our blueprints would be 100% identical, I would see this as a problem because I wouldn't feel me as I feel now, right?

RANDAL _ Right, but since I'm over here saying different things feeling different things we already know that we wouldn't have the identical blueprints, even in twins it wouldn't be the same. So, you know that isn't going to happen. The only time that is going to happen is in fact at the moment where you make a copy of an upload and at that very moment before they start experiencing different things they are in fact completely identical. But it means nothing because before they experienced things time hasn't started to run for them, so they can't sense anything, they can't be aware of anything, they can't feel that there is a problem.

FILIFE _ OK, that's a sort of Schrödinger approach, a quantum approach right?

RANDAL _ Well there's no quantum in that. You have an emulation, if the emulation is not running if it's not generating anything not producing anything then cognition, because they were just made a copy of that emulation but you haven't had ... no time has gone by, yes they are identical but this is something I try to explain so often whenever people talk about things like identity and consciousness, often it's talked about like this thing that you can have in a state, where a state is something what you are, a specific point in time. But in reality, you're talking about functions and things that produce an output over time. So, time has to go by to experience something to be aware of something to feel something. So only when these emulations are then progressing through time does the brain generate any experience. Does the experience have a chance to form a way of wandering "w a i t t h e r e ' s a c o p y o f m e"... and while that is happening

the two of you are already beginning to experience slightly different things and you're not quite the same although you are very very closely related.

FILIPPE _ OK. So, I guess we already discussed the last question I have here...

QUESTION V

FILIPPE _ Is it inevitably going to generate a dystopian society?

RANDAL _ The problem with a dystopian society is how would you define dystopian? I think that for someone who believes that we should never temper with our biology and that we should live as we did a hundred years ago or something like that, say you're *amisch*, then yes perhaps this is a dystopia that is coming up. But if you would live at some point where we could live beyond Pluto or something then it's not a dystopian society and that is what is necessary to be able to get there.

FILIPPE _ Sure, but if we get to the point where we can actually split people into those who have access to enhancements and those who don't, we certainly...

RANDAL _ I see, so but then you're imagining a setup where there are two categories and you are discriminating between them in some sense that we do now say with chimpanzees or something like that... But what I imagine is going to happen and that is just a different perspective that we have is I imagine that as soon as we can actually experiment with our cells with our development if we're not going to have two groups at all it's like a gigantic tree of life like a cambrian explosion of possibilities. So you're not just going to have the uploads and the biological humans. The biological human can choose to be an amish farmer and you're going to have the somewhat enhanced biological person who wasn't comfortable with uploading but likes to have brain enhancements and you're going to have the person who has specialized in interstellar travel and don't have a body at all, they are a spaceship and have lots of these little drones in order to reach a planet or something and you're going to have a zillion other possibilities of what a person is. And so then you just have to expand your notion of who belongs into that society of sentient beings, the ones that we give personhood to. And that always has been something that has expanded, you know it used to be that we only cared about our tribe and outside of that if we'd kill someone no big deal and then you started to care a bit more and then at some point we decided well maybe women are

people to and decided to give them the vote... and at some point you start to wander maybe we should extend personhood to artificial intelligent agents over there or to this enhanced person and this upload and you get this bigger circle, really.

FILIPE_ OK. Last question or sub question until what extent are these enhancements leading to life extension? You align with Aubrey De Grey, he says that in about a few years, well two or three decades or so, you could live up to a thousand years (1000).

RANDAL _ Oh ok that's what you mean about "until what extent", ok. So I'm not sure if you're asking me when could we live...

FILIPE _ No, to what extent, I mean you think it's possible for us to extend life up to a thousand years just as Aubrey de Grey says, or even more if we could really eternalize ourselves via MU...

RANDAL _ Ah ok, so my answer is I do believe it's possible. I don't often spend a lot of time thinking about that because as you may have noticed in the conversation, I really see the most important aspect of this being the adaptability that we gain and the way that we can then explore it into many possibilities as an advanced species. And so some of those in that species will have a very long lifespan that is just one of those possible directions of development. Some will choose to have their identity unmixed with others and simply who they are kind of like how we experience ourselves going forward for a thousand years. Sure some will go that route. And some may even go for ten thousand without diluting into something where you wonder, wait is that still the same person? or has it become something completely different? You begin to ask those questions after a while, just like am I still the same person I was when I was four years old but when you're a million years old, how much of you are you still going to be if you would live that long? But, yeah, sure life extension beyond what we live right now is possible.

FILIPE _ OK. I guess that's it. These were the questions. I want to thank you again. I think you have dutch ancestry, right?

RANDAL _ Yes I do. Dutch and more. My mother is German from an austrian background and there's a bunch of polish from another ancestor too, so it gets crazy. But, yeah, my mother tongue besides English is Dutch and German.

FILIPE _ Then I would say "Dankjewel" in Dutch and "Danke Schoen" in German!

RANDAL _ *mai pleisiere*. Thank you! It was a pleasure!

FILIPPE _ No, it was mine! One last question: would there be someone you'd advise me to specifically interview?

RANDAL _ You mean specifically in Mind Uploading?

FILIPPE _ Yes MU.

RANDAL _ Ok one person you might want to talk to is Ken Hayworth

FILIPPE _ Keith Wiley?

RANDAL _ Well Keith is also a person you could talk to. He has plenty to say.

FILIPPE _ Yes, he hasn't responded to me yet.

RANDAL _ Ken Hayworth is a member of The Brain Preservation Foundation. He's definitely worth talking to.

FILIPPE _ Ken? Yes, I watched him too at the **carboncopies** webinars. Ok. Thanks again, are you planning to publish a book sometime soon?

RANDAL _ I'm always planning and haven't yet got around to it, yeah.

FILIPPE _ OK. I'm expecting that ;)

RANDAL _ Ok I'll stick to my promise and try to write one! So have a great day!

FILIPPE _ OK. You too! Thanks. Bye bye. **THE END**

APPENDIX E _ INTERVIEW WITH DAVID CHALMERS

(Philosopher / Cognitive Scientist / Adelaide University, AUS)

(answers sent by e-mail @ MAR2020)

1_Regarding Whole Brain Emulation and mind-uploading

Q1: Would another me be me? (or just a copy of myself)

They'd be two different people. Both might have equal claims to being identical to the prior person.

If so, then How would you know you are you?

If not, then..... What would our copies *really* be?

Q2: What would represent the biggest risk in achieving something like this?

2_Regarding Superintelligence and the Technological Singularity

Q1: If (When) it happens (*and we're still humans*) will we even know it (*be aware of*)?

Yes, it will be obvious.

If so, then Could it be because we would be *less bio and more bit* at this specific moment in time?

If not, then..... Would it be because we would have failed to cope with progress?

Q2: How possible would it be for a *Singleton* of some kind to rule the World?

Quite possible.

3_Regarding Transhumanism and Posthumanism

Q1: Will *being human* be different in a Post-human Era than it is today?

Almost certainly.

If so, then Would this mean we would definitely be *more bit than bio*?

That's likely eventually.

If not, then..... Would this mean that an eventual cyber-sapiens of some kind would in fact be the next natural step in human's evolution? **This too.**

Q2: Do you think politics in a few years will be divided between bioconservatives vs transhumanists?

There will certainly be such a division, but I have no idea how important it will be in politics (probably important at some point, at least for a while).

4_Regarding Superintelligence and mind-uploading

Q1: Would it make sense to iterate human intelligence via mind-uploading to augment it / improve it, and in this manner cope with AI's progress? **Yes.**

If so, then..... Could we avoid a singleton of some kind or AI's progress would sooner or later always develop into a *Superintelligence*? **Maybe, but there is still a danger of singleton augmentation.**

If not, then..... Would *universal darwinism* not select AGI as the *more adapted intelligence instead of ours*?

Yes, though I'm not sure that putting this in terms of Darwinism is helpful.

Q2: *Could mind uploading prove that two different persons sharing an identical blueprint would in fact be identical?* And if so, would this not shake our beliefs in issues such as identity and personality?

No, I don't see how it could prove this.

5_Regarding Transcending Biology and Human Enhancements

Q1: Is it inevitably going to generate a dystopian society?

Not inevitably, but possibly.

If so, then Could/Should we trust our fate in a system run by an AGI?

It depends on the AGI. (The AGI may be us.)

Q2: Until what extent are enhancements ultimately leading life extension?

Once we have uploading, relatively immortality is a consequence.

(Computational Neuroscientist / Future of Humanity Institute / Oxford University, UK) [skype@11MAR2020~60min](#)



FILIPPE _ Regarding Whole Brain Emulation and Mind Uploading, would another me be me? Or just a copy of myself?

QUESTION I

ANDERS _ So I think the issue here is of course of personal identity as well as a related question can you have a conscious mind existing in a software which is a bit of a separate question. Now, I take a functionalist perspective on philosophy of minds, I'm actually fairly confident that yes, you could probably have software minds. So that's not the problem. I also don't think personal identity is a problem because I don't think personal identity exists! I actually use Derek Parfit as my landlord and I guess some of his use has rubbed it off on me. So basically when I say I did the following things I went to school, that is of course because I have memories of going to school and that personal existence at school has a lot of similarities to me and I may endorse that 7 year old Anders is me in some sense. Similarly I might talk of 70 year old Anders hypothetically in the future but this is a psychological continuity in some sense even narrative one. It's not, I might have a religious revelation or change my opinions in the future and at that point we might get a really interesting question: would it now endorse that religious Anders as me? And if it's a sufficiently big break I might actually say well that's a different person. So when we get to brain emulation, if it really goes very smoothly, we might imagine a small example where we do it continuously, It seems like continuity has no problem. And I think it prompts our intuition to say yes that's going to be a continuation of ourselves, our narrative is not broken. In a more kind of freeze slice and dice perspective that I think is more likely for at least early brain emulation. There is going to be some dramatic shifts here in how you exist. First you are a living person then you undergoing this process and then you are a software person. And then at that point some people would probably say yeah this is enough of a break for it to not to be me. And others might say No, this fits with my sense of identity. I'm fine with it. Since I regard myself as a ?!... sufficiently like I'm fine with this as you mean it doesn't disrupt things too much and indeed I'm very happy with multiple instantiations of me they would probably say yeah we're all Anders.

FILIPPE _ OK. Regarding this what would be the most dangerous risk in achieving something like MU?

ANDERS _ Well, there is some fundamental scientific and philosophical or technical reason why it cannot be done. So I think the most clear thing in science that I would like to investigate and I think we should be pushing for is: is there enough scale separation in the brain to make it possible to emulate a brain properly. So my standard example is when you have a gas you can average what the molecules are doing and talk about the pressure and the density and volume of the gas and you don't need to talk about the details of the gas you can actually model the gas very accurately without thinking about the small stuff. This is not true for a turbulent flow of water. Because there the microstructure affects the mesostructure that affects the macrostructures. If you don't simulate it on all the scales you get what a computational fluid dynamics people call dry water. It's like water but it's not water. Now, does scale separation happen in the brain? I'm kind of thinking it does. I do believe that we do get a fabric once we get down to the subcellular model, I'm kind of following the standard computational neuroscientists assumption: Yeah you can divide the neurons to isolate compartments and that's well enough. Inside the things are mixed because the brownian motion is effective over those distances. So at some point you have scale separation. But this remains to be proven, and if it's not true, in that case you might need to simulate things on a finer degree. So we'll eventually run into hardware computational problems. So I'm fairly confident that we don't use quantum computation and in our thinking like the physicists like to bring it. But if they were right about this then it might be a problem in principle to copy things. So that's kind of the main thing that I think would be hard to do. Then we have the philosophical issues. Maybe Searl is right and computers can't actually have conscious states at which point other really weird things happen on a brain emulation because you might get a philosophical Sanders for example. I don't think that's likely but then again philosophy of mind is hard and I don't think we understand consciousness so we might find out something surprising. So I think that are the main reasons why it might fail. Of course, we have other reasons why we may not succeed as a research project.

FILIPE _ So you think WBE could fail?

ANDERS _ Yes.

FILIPE _ But you know that at the moment they've already achieved 90% WBE in Mice.

ANDERS _ Not really. So it's one thing that I can take my super computer and run a very nice set of Hodgkin-Huxley equations for a big piece of simulated cortex and get roughly

a course of activity. That's not the same thing as WBE. What you really should do is you take a mouse that you trained to run a particular maze. You scan its brain and now simulate what it does and see if it runs in the maze the same way. But it knows the cheese is on the left and not the right side. That would actually be a real brain emulation and would demonstrate we've learned something. That's promising but it's not WBE. And we haven't achieved that with the mouse at all. We've done small models of the cortex. The Henry Markram Blue Brain Project, they all demonstrate beautiful stuff that looks biological but we can't test it yet. So we're moving in that direction but we actually haven't got the right proof yet.

FILIPE _ OK. So you don't believe that much in the actual possibility of MU?

ANDERS _ I believe, here's the fun thing, from your voice it seems that I am a very big skeptic. Of course most people would say ...

FILIPE _ Well I'm a skeptic myself also because of my scientific background I tend to be skeptical. It's a default program in my hardware or software...

ANDERS _ Yeah. Well most projects fail because of various reasons. The question is to find the reason is fundamental or funding run out or we didn't find somebody that could program the thing we needed to program. I think it's very unlikely but if WBE never happens it's going to be because of something foundational. But I do think it's going to be a very tough project to actually make a working WBE. I think it's worthwhile and I think it's something valuable to put effort into doing. But the funny thing is, among my transhumanist friends I sound like an extreme skeptic because I don't think it's guaranteed to happen. That of course many other people would say wow you think that there's even a chance that it happens that is mindblowing. So this is why I have this kind of position in the middle.

FILIPE _ OK. Let's move on to the second question. Regarding Superintelligence and the eventual Technological Singularity, if and when it happens (and we're still humans) would we even understand it? (be aware of it?). AGI could develop by itself in a few seconds and in that timeframe we wouldn't even understand what is happening.

QUESTION II

ANDERS _ So I had this discussion with Nick Bostrom about this when he was writing his book of Superintelligence. Basically what is the speed of take-off. And the way of measuring the speed is to think about it relative to human reactions. So you could have a hard take-off but it's false to compare to human reaction times. Or you could have a slow one where our society and the institutions can change at the pace when it is happening. And then you can imagine some middle ground where it's happening rapidly but it's still time where you can make some decisions. Now, I don't think we have actually any good evidence for whether any of these three is most likely. So one one hand there are interesting arguments like Robin Hanson's economical arguments pointing out a hard technological singularity take-off but it's a kind of very weird claim about economics basically saying that a subsystem can become the whole and grow much much faster than any other part of the economy. That is a weird claim to an economist. They would not normally believe that. And it is kind of a weird thing because the systems of course always try to be better, companies always try to be better than the economy. You could make the same argument in ecology or biology and say could a small part of the biosphere take-off suddenly and change geography? And at this point, the biologist might look around at humans and say oh actually it seems to be happening right now. Because from an evolutionary standpoint what has been going on is up until recently the biosphere was kind of evolving at a normal pace and then some species goes around the whole thing and changes the rules for evolution. That is kind of an evidence but you can't really make something relatively fast. Unfortunately neither of these tells us something about time-scale. My personal guess is that it's probably not going to be superfast. There might be problems about gathering evidence and data to be effective at being intelligent. But then again it might be much faster than human reactions. Another issue of course is how visible this is? So in a more distributed cell you could imagine something that would transform the world and society and things would get very strange very rapidly. We could also say now we have a piece of software that is outrageously intelligent and we don't even notice how things are going on just as mice can't notice the clever plans humans are making around them. So I don't think we have any good tools to analyze this. I think what we can do is analyze the different scenarios and try to see if we can do something sensible about this and then try to take the right kind of action. For example, Eric Drexler has a critique about the AGI idea in these comprehensive AI services and points out that actually AGI is probably not what most of the companies would like to develop. The

business case for AGI is very weak. But you could imagine something like the Amazon Cloud for intelligence. It should be useful if you could connect various sources of intelligence and skills to various tasks. You could make lots of money and indeed this is much similar how we normally do in engineering. So in this very big wide paper he sketches out a pathway of the direction that seems to be very much likely then in that scenario you don't end up with a technological singularity in the sense of one little server somewhere waking up and taking over but rather that suddenly you get intelligence on tap. We get as much intelligence in the world as we connect data servers to the cloud. And of course we can use these servers to come up with better ways of making better servers and better technology. So you get a formal augmentation here which might very well mean that some people would augment themselves etc rapidly in an extremely strange situation. But it's not based on any particular single superintelligence existing, it's rather superintelligence as a service. Now this is a very different scenario. Of course, your other here is about the singleton issue, which I think is a very important one because of the AGI scenario. That is relatively easy to turn into a scary story about how things can go badly wrong because the AGI is very powerful and it might have random values. The AI servers scenario on the other hand could also go very badly wrong because you can imagine evil people using superintelligence as a service to support whatever their evil plans are and there are lots of them. On the other hand it might be that all superintelligences are owned by J.F.Besus and his wife. So it's all in the amazon cloud. Now, when Nick wrote his singleton paper ..

FILIPPE _ You think that we'd all be part of this singleton somehow? Not a unique computerized entity but rather all together in the cloud?

ANDERS _ That's a different meaning. Now you're talking about a group-mind. Nick Bostrom's definition of a singleton is about agency. Who decides what's going on in our civilization. So right now we don't have a singleton. We might have various kinds of corporate entities where we have group-minds like us, the voters and various organizations. And you could imagine a world run by the United Nations ++ (plus plus).

FILIPPE _ Or google.

ANDERS _ Or google!. Now, the interesting scenario here is: Suppose google could decide OK - we don't want anybody to work on biotechnology that is too risky, we should shut down all biotechnological development. In this hypothetical world they could actually

do that because they have enough power to do that, similarly the United Nations ++ or maybe the big AGI. In that case you would have a single point of agency that could actually determine a large scale of things going on in the world and it would also be capable enough to say: Yeah, you tried to stop it but I can stop you from stopping it. Now, this is an issue of power and control rather than we all joined up in the cloud. Now, you could imagine getting a singleton because we'd all join together in one vast group-mind, and this group-mind is making decisions for all mankind because it literally is mankind. That's another form of singleton, which is interesting because in some sense it is legitimate. We might have trouble with Google running the world because who gave them that job? Well they had the first AGI was just sheer luck. But if we are all part of the group-mind we might say: ok, yeah, I'm running the world so I'm totally fine with that. Now, the problem is of course is this a likely scenario or not? I think, again, we need to analyze different possible scenarios. On one hand we might imagine that we see various risks with technology or various economic developments that happen that lead but the current ??stistic of the world disappears. So this is a bit of a theme that goes on in Nick's recent paper about vulnerable world hypothesis. So he's basically showing that there are potentially technologies or treats out there but we probably need to figure out a way to get our act together. The current mismanagement of covid-19 is a pretty good example. Inside China, its government is almost a kind of singleton and actually decides what happens. And it's an interesting case because they failed at first. They did stupid things early on which probably lost them the whole game because they kind of didn't listen to the doctor etc... And then they did the right thing and shut things down, and they are now controlling the virus well except that since it leaked out it doesn't matter, now they're going to be reinfected by the Italians and so on... But we could imagine a world where we have dangerous technologies around and maybe we just need to have a system that if something dangerous shows up all the governments need to join forces in an effective way to do that, which of course sounds very utopian by current standards but might be something we can build the tools to build the tools for. At this point many people are going to be worried about how to be sure that this doesn't go badly wrong. A singleton is an existential risk too, if it has the wrong values. If you have your good AI which is running. But in any case the singleton scenarios are interesting but require that one develops it. You can imagine other forms of multilateral scenarios where maybe the US and China develop an AGI together with a joint product to at least get a safe AGI. You can do all sorts of very different things here, it's not like it's a binary set of options.

FILIPPE _ OK. Moving on to the third question, regarding transhumanism and posthumanism... With all these enhancements we're getting and shortly get even more in the next ten twenty years, with exponential progress and even if we account for things like MU, would you think we could in some way direct into the next natural step (in *italic*), into some kind of cyber sapiens?

QUESTION III

ANDERS _ What does *italic* 'natural' mean to you?

FILIPPE _ This is tricky. I don't think that evolution would naturally evolve in that direction but we'd naturally delve more and more into the technology and it would make us some kind of cyber sapiens *naturally*...

ANDERS _ I think there's a lot of fun stuff to unpack here. One is of course what we mean by natural. We are of course applying what I would say a kind of industrialized or post-industrialized era concept of nature to it. If you think of hunting gatherers they rarely have a concept of nature. They're not separate from nature, they're living in the middle of it. When they refer to nature they're just referring to the state of the world as it is. Then you get to the agricultural era when people are making their own cultured field, and now we have to do a separation between the wild and the cultured. And the wild is very threatening and very problematic because it might overwhelm the cultured thing because of survival. So now nature becomes something quite threatening. And in the industrial era of course, besides that we might at this point when we become more environmentally aware suddenly nature becomes this other thing that is precious and fine and we want to go on holidays and see it and we want to preserve it and we kind of feel that it is in some senses weak and fragile. So we have been changing that. At the same time there is this important aspect of where is the route of your life, where is the foundation of existence. Many people definitely want to have a sense of good foundation for their own individual existence and in some cases you might say oh I recognise that I'm a member of the species, I'm a member of the biosphere that gives me a sense of being part of something bigger than me that makes me very happy. Other people might of course use this to talk about their nation or other social organization. So nature also becomes a point of allegiance? Now the interesting part is that you can interpret this in different ways because many people use basically the great chain of being, that medieval concept of

well on Top we got God then humans then animals, then plants, then organs then minerals. There is kind of an organization, a structure of the world. And upsetting the structure is not just changing things. It's actually morally bad. And this is why I get a lot of resistance to ideas that change nature and of course especially transhumanism and posthumanism. Now others would say: oh wait a minute, evolution is ongoing here ,we are doing the natural next step here by modifying ourselves. By natural next step what they mean is they're trying to use natural here as a good word to depict that it's part of the natural order, which is usually implicit. But you go back to people that are around me and this rational dignity of man, which is important for transhumanism. In that document he argues that God gave all the animals different properties; Lions are brave and fish can breathe underwater and humans got this fantastic ability to change themselves. And that is what gives us dignity. And that is what we usually do to get a concept of modern human dignity. I'm often briefing the case of human dignity to transhumanists and it's very fun because I've been arguing this with Francis Fukuyama that of course has a much more aristotelian idea about human nature. No, modifying yourself means going outside the human, it can't be a good action for a human to become not human hence transhumanism is bad. It's very fun to have these arguments. But getting back to the question here I think there is a descriptive aspect of nature and evolution and where this allegiance aspect of tribe and a moral one.

Evolution is actually an interesting process in that it invents new ways of evolving. So the original evolution on Earth really got radically changed when sexual reproduction was invented by evolution, accidentally but still suddenly.

FILIPE _ In fact 3.8 billion years ago by bacteria.

ANDERS _ Yeah. So evolution has come up with various manners that change how evolution works. So for example mothers give some of their antibodies to their infants through breast milk. This is in some sense transmitting information between generations using a non-genetic factor, getting around the bottleneck of genetic evolution. And then of course nervous systems with learning abilities get around another problem about updating neural connections and then a while ago human brains which are able to transmit information by talking.

FILIPE _ Yeah, and now we have technology. As the next step...

ANDERS _ Yes. So the interesting thing here is where do we say that natural evolution has stopped. I would say no, actually natural evolution is kind of discriptional how new forms evolve. It's just how the mediums get changed quite a bit. So by this account I think getting to cyber sapiens in a post human era is totally natural from a descriptive standpoint. It's just what I would probably expect if I had a million intelligent creatures evolving on a different planet. I would probably say that I expect most of them if they develop this kind of technological civilization to do something like this. Then we might have different goals and styles. So the other thing is of course, well the tribal allegiance. How different would you actually feel if you were a post human. And that might be really complicated because there is one thing how you feel about your body and about your life. And then what story you tell yourself about that. So if you think about brain emulation, in some sense it's the simplest way of making a posthuman you can imagine. You take a brain, scan it and you run it in a computer. And by assumption if this works you get the same mental states that you would have had with the original person. If you take Searl and simulate him as a brain emulation you will get a philosopher who is going to have some philosophical trouble with the fact that he is conscious, assuming he is conscious. But he would still have the same beliefs and the kind of human aspects. It's just that he happens to be part of the mineral kingdom now rather than the animal kingdom, except that if this has worked really well, simulated Searl would still feel that he is the original body. That might of course change as virtual reality and his avatar, etc... Similarly if I wanted to add various extensions and enhancements to my body in some cases I might want them to feel very different from how it normally feels, because that's the point. I want to deliberately extend myself because I want to feel transhuman. But probably a lot of utility enhancement that I don't want to feel at all. I don't want to get sick and don't want my vaccinations and other things to make me feel very different, they should just fix that. So there is an interplay here in their living experience. And this of course gets even more complicated because we are unfortunately being social creatures rather likely to create various tribes and ways to organize it. Sometimes I fear that we might split off into the windows and apple civilizations. We might have simple people that do upgrades but are not compatible for trivial and technical reasons. But then they become design reasons and that affects the way of your life and then of course both of them are going to have trouble in understanding those weird open-source Linux minds out there. So, what I'm hoping of course is that we get a lot of diversity. We can still hold together in a joint society because we have enough in common. There are always some people who

want to isolate themselves and say we here are particularly particular for some reason, and that makes us better than others and that's going to cause all sorts of friction.

FILIPPE _ Yes, let me take those words because that could also lead to a dystopia in the next years where there will be those who want to be enhanced and of course those who don't. Do you think politics would in a few years be divided between bio conservatives and transhumanists?

ANDERS _ Yes. Just look at many of the european green parties. They are left-wing bioconservatives. Many of them have a very strong idea about the proper ecosystem and implicitly also about what the proper social system is, and they actually have way more in common with evangelical christians in america that also got strong views about what genders must exist etc... And the issue however, while I could imagine that the future political debate between transhumanists and bioconservatives, I think it's not going to be that nite. After all, why should environmentalists be on the political left wing? Indeed historically they actually began on the very conservative right wing and it's kind of a weird coincidence that today you find most of the environmentalist movements relatively political left. And this has to do with compatibility between different ideas and also political alliances people make. Libertarians, they made a Deal with the conservatives in many countries. A kind of joint conservative alliance which meant that they had to share parties with religious conservatives who are actually anti-libertarian. That worked in the nineties and then started breaking up more and more and more as populists showed up. Today I can totally imagine an alliance between the cosmopolitan-internationalists and transhumanists against naturalist-bioconservatives and environmentalists. You might see all sorts of these alliances. You might imagine that maybe the green party takes over from the inside by the bright greens, who are pro-technology and think that oh climate change is horrible: we need more nuclear power. That is the kind of green party we're seeing in some places and of course that's going to cause a lot of trouble with the more traditional kind of dark-greens who think "No no we should go for a small society instead". No, you can't have the global problems with small scale societies. We need large scale smart societies, the bright greens say. I think it's going to be interesting but very annoying because you do find some of your friends' views on transhumanists sometimes on the opposing side. I'm very disturbed and I come across a neoreactionary transhumanist. I find that downright disturbing. Of course they also got rather stupid views and keep rather silent when Trump got elected, because he's a total counter expample for everything

they were claiming online before: oh having a businessman running the US like a business is going to be great! HA! (HIHIHIHI). But I think the real issue is of course shaping the view, well what is the world like? What are people like? What do we do to give these views? And in this point transhumanism has an advantage in many ways because it's forward looking. Quite a lot of current political ideologies are very much based on past conflicts or past ideas. Basically the traditional left-right argument has devolved into a question of how to move money around. The conservative vs liberal argument is more alive actually. Do we actually want to have social structures that are in some sense not defined by ourselves or do we get to define our structures in a flexible way. That also is linked to the nationalist globalist perspective but it's entirely possible to imagine transhumanists ending up on all sides because transhumanism as a movement of ideas has a weakness. It doesn't have a theory of a value internal to itself. So you can be a christian transhumanist. That's no problem. You can be a fascist transhumanist. You can be a libertarian transhumanist and still plug in your own value of theory but of course christian and atheist transhumanists have a little disagreement. The fascist and libertarian transhumanists definitely have a big disagreement. So the real question may actually be how much politics is going to be driven by this and how much is going to happen also with other non-politic players like cultural technologists.

FILIPE _ I'm guessing a lot, because once you get these differences in terms of who gets these enhancements, who has the right to have them, who's going to govern this..

ANDERS _ And different transhumanists give relatively different routes. Most transhumanists that you would find today of course have a fairly liberal view : oh, morphological freedom is very important, you need to be able to control your own body, others should not be able to control how you upgrade yourself etc... Many people think that is very problematic but then they get very nervous about the alternatives, so the government has to find what your body's like, that sounds even worse. Of course, bioconservatives, their solution, if we just avoid developing these technologies, then would be fine, because then the question doesn't come up, which of course misses the point that these technologies already exist. People in my circle of friends are slowing down ageing by medical, others have done gene therapy to deal with it and some others are tooling around modifying their brains. So the bioconservatives can't win that one.

FILIPE _ Yeah, but probably I would say that the rich will get enhanced first and the poor eventually die out.

ANDERS _ Why? Why haven't the poor died out already?

FILIPE _ Well you know because socialism cares for ...

ANDERS _ Socialism actually killed much more poor people than capitalism. The real story here is new technology spreads unevenly. But when the price of a technology comes down, gadgets like my smartphone are super advanced ...

FILIPE _ Yes, it's "democratic"

ANDERS _ Hmmmm, it's not even democratic, it's about technology diffusion. When Apple made the i-phone it was really expensive. Did they try to sell it just to the rich? No, that would have been stupid because there are not that many billionaires to sell it to. Selling to the upper-middle class on the other hand, now you can make a lot of money. But of course as smart-phones spread people developed cheaper smartphones and soon everybody found use for them and today more than half of humanity has a phone and practically all of them are pretty smart although they're not i-phones. So what happened was the price came down exponentially and became democratized by diffusion. Now the early adopters could (those that had really smartphones), but today that's not something you want to boast about. If you want to boast about your smartphone today you need to have probably gold plated with diamonds or something ridiculous. Now, the same thing for drugs. The price of drugs can come down exponentially once patents expire. Services on the other hand tend to be really expensive. Why? Well if you need to go to an enhancement spot to get upgraded you need to pay for the doctors and the nurses salary and of course your own time-out from job. That also is costly. So enhancements are kind of service based. They are more problematic from an equality standpoint from the gadgets. I think we have an intuition here that maybe upgrading your body will be more like a service than a gadget. That is an interesting question which technically one can properly analyse separately for different forms of enhancement. I think the goal for many would be to turn many services into gadgets, you know swallow a pill or logon to the right web site. So when it comes to where power resides in society a lot of it has to do with where does the benefit go. So what we've seen over the past twenty years is that having a great education has been super valuable. Because we, with a higher academic education, we have been able to use the new technologies and we're not that enhanced, not drinking better coffee or taking much more modathanyl than we've used to but we have a lot of technologies that actually amplify the abilities of smart people

to use things. And that has very little to do with people themselves or with super enhancement. It's just that they fit well to the technology. You could imagine a scenario where something similar happens with enhancements technologies. On the other hand, many governments would dearly want to enhance most of their citizens. Just imagine if you could boost the immune systems of your citizens. They have less sick leaves, they are more productive and don't get hit by pandemics and they're probably going to thank you for having a great health care system. Indeed many forms of enhancements might be just the thing governments would love to subsidize. If it turns out that this is good for people and the economy then they would want to do it. Which is of course why governments are subsidizing education rather than saying oh no let's just have billionaires run the universities. Instead we actually use tax pay money and make schools mandatory, which I think is really an interesting example where we made one kind of enhancement into a mandatory thing. You can't be part of our society unless you can read and write. So I think we are going to see these debates showing up again because it might very well be that in fifty years that governments are going to be very concerned if parents want to have kids that don't have genetic upgrades. If only for the sake of the kids. And a lot of people will say those parents are kind of harming their kids in a weird way. Now, this is a bit like the social shift. Today you probably can't live a life very well without a smartphone. My mother doesn't have a smartphone. She doesn't even use a computer. And she's getting more and more upset about the fact that she can't pay with cash in many places. So, she doesn't want to use the card. She wants to pay everything with cash and everybody else tells her yeah, Mrs Sandberg that would have been nice twenty years ago but no longer, and my grandmother wanted to pay with cheques and that was kind of hopeless already in the ninetys. So we have this interesting thing that we get enforcement not just in the sense that the rich leads. But also enforcement in how we set up our society in large. And that enforcement is in many ways much more powerful. And sometimes very disturbing because in a society that is cashless how do you give money to a beggar? How do you give money to a kid? Well in Sweden beggars and kids have swiss?! accounts you can use them with a phone of course, but that presupposes that everybody has a phone.

FILIPPE _ Yes, we kind of covered the 5th question. Not the 4th. Because we were talking about the possibility of the dystopian society and I was going to ask you if it is inevitably going to generate a dystopian society...

ANDERS _ By the way, you should notice something interesting here. We have been talking about inequality and I think you are assuming inequality is what gives you a dystopian society. But a lot of great literature dystopias have been about equal societies. If you think about XX?!byYY? That's totally equal; If you think about 1984, there are some very nasty inequalities but within each stratum of society they are extremely equal; And The Brave New World again, there's actually a weird kind of equality going on. The reason of course they are dystopias is not just inequality. Imagine the Star Trek society where everybody has everything they want for free but exists one zillionaire owning half of the galaxy. Would that suddenly make the Star Trek society bad? And it seems that, no, that's not a problem. The problem in the brave new world is that there is no real creativity, no real human spirit, it's kind of whenever you try to think uncomfortable deep thoughts somebody gives you drugs; In 1984 of course it's a totally oppressive society. And in "XX"?!, It's the individual that is crushed. And you can get all sorts of really weird disturbing dystopias you see in enhancements. For example Nick Bostrom in his article "The Future of Humanity", he brings up one where you have mindless outsourcers. So we outsource more and more of our brains to use for apps in the cloud and eventually there's nothing left. So you have an economy that is growing tremendously and superskilled entities doing stuff but there's no consciousness there's nothing and nobody at home! He describes this as Disneyland without children. Now this is an interesting dystopia because it's bad, but it's not even bad to anybody. It's just a waste of potential. So I don't think it's inevitable that we end up in dystopias. I think it's inevitable that whatever the world turns out like, somebody is going to say this is dystopia. But I actually don't think that the world has necessarily become a dystopia that we all agree is a dystopia. I think we have a good chance of getting tremendously utopia worlds. Just if we think about the current world, it's this mixture of horrific dystopia and beautiful utopia at the same time. All knowledge of mankind is more or less freely available to anybody and by the way we have oppression using facial-recognition in the interests of china it's this crazy mixture of very good and bad and I expectact this to continue so we will probably be horrified and delighted at the same time by the future whatever it turns out to be. Unless it goes down in one of the straightforward dystopias. Yeah, sorry, a little round there..

FILIPE _ No no no, great answer, yes! So the last question. Regarding SI and MU, would it make sense to iterate human intelligence via MU to augment it and improve it and in this manner cope with AI's progress?

QUESTION IV

FILIPPE _ What I mean exactly by this is I have a kind of a work in progress of a thought experiment of a scenario where you'd get a substrate independent mind, as Koene describes them, so a SIM1 that you would then compare with lots of other copies SIMs of your own mind, and you would then run them (like any kind of file) in different environments and competing and so in this manner you could select in a darwinian kind of way you could select your best answer (like in chess), and foresee your answer, or more serious environments, like the coronavirus situation, you could foresee your best answer after iterating your SIMs and you could edit it (as you can edit anything digital), and change it with AI and make it a little bit different of SIM99...

ANDERS _ So I think this makes sense, yes I quite often think that it would be quite useful and be able to try modifications if we exist as a brain emulation. You might change some receptor potentials or other things, and then after a while you test whether this was good and if it wasn't good you can always undo the change and revert to the original. Of course you can do it like you described and have these kinds of parallel evolution. It might be that you'd tried out a hundred versions of yourself and see which variation helps the most and then you'd keep that one! Now, a lot of it depends on your view about what you do about spurs, these small copies that might not have a long future. Have you read Robin Hanson's "The Age of Em"?

FILIPPE _ Yes.

ANDERS _ So he of course thinks that most people would be kind of happy to have a lot of spurs and delete the spurs. I don't think that's true for most people. I think in theory I'm ok with it but I'm not certain I would actually be ok with it if I was in that situation, but it's worth considering. There is this issue of how much are you allowing yourself to use versions of yourself as a resource. So Nick Bostrom has this concept of Mindcrime that shows up in his book. But it might be that you get bad outcomes simply by having a lot of minds in bad states going on inside some bigger system. You could imagine a superintelligence with vivid characters but the characters have lives and experiences and if we think of most characters in most novels they actually have a lot of anguish and it might actually be a very bad thing for them to exist. So there are ethical issues around it, but I do think it's potentially quite powerful in making new modifications especially if you have opaque systems. Brain emulations are essentially black boxes. It's a model of

the brain and it has all the messiness and we're probably going to learn a lot of neuroscience once we have brain emulation because we can do a lot of experiments including neuroscientists that can do experiments on themselves and undo their effect, and then report back on what happened. So this is powerful for black box systems. If you have AGI based in some sense of understandable code and you can read the code and prove things you get a very different form of improvement path. The question is which one is going to win and I have a feeling that it might be easier to improve AGI than human brains because we have a backwards compatibility issue which is pretty tough. I think combining human brains with AGI will be very powerful but now you need to think they are very very different systems and that's going to be a tricky task. So I have a feeling that in the long term, I have my doubts that humans are going to compete even if you try to upgrade them. I might be wrong, it might also be like Robin Hanson's suggestion and human like minds might be part of AGI systems. So in fact it might be that you and a few others form a team that is actually essentially the brain stem of an AGI. And you're going to be feeling that you have the best job in the world and the AGI will think "hahahha" I'm in charge my brainstem doesn't control me as our brain stems totally know that they're in charge.

FILIPPE _ Ok, now just to finish and still in this kind of scenario, imagine just as you can compare DNA sequences as blueprints, and you have yours and I mine, imagine you would compare yours with mine and they would be 100% identical to each other. Would this shake our beliefs in personal identity? Until what extent could I claim to be a different person?

ANDERS _ Yes. I think that, and getting back to a historical conversation here, I'm with Parfit. This shows that personhood and personal identity isn't a real thing. I don't know if you read his book "Reasons and Persons"?

FILIPPE _ No.

ANDERS _ It's a classic, The nice thing is, ignore the "reasons" part, the first third, because that is really incomprehensible in ethics and ethical truth). The "Person's" part is really fun, because he's starting to use the teletransporter. He was the guy who came up with the idea about personal identity in this Star Trek like teletransport. Funny thing I found out this week, I was having this chat with Meselsohn (geneticist), he was the guy who proves that DNA replicates by the two strands separating and kind of copying a new

strand. Turns out that when he was a grad student back in the fifties he was in a philosophy debate with Richard Feynman about personal identity and they essentially came up with exactly the same thought experiment about the teletransportation part. Both Meselsson and Parfit, their idea was ok this assembles you and reassembles you. It uses different atoms etc. Oh you accidentally end up with two copies. But then Parfit goes on, because he's a philosopher, it's his job to go on. And he actually uses a bunch of examples where you merge people in various ways. So you end up with two people with very similar or identical connectivities. There is also a very good fictional story based on this idea by Greg Egan?? called "close" where our two post-humans also do this teletransport in their slightly weird marriage. Now I think Personal Identity doesn't exist like we normally do think. Our concept of personal identity makes a lot of sense in everyday life just like free will makes a lot of sense on a human scale. I don't think free will makes any sense on a neural scale or atomic scale. I think I'm a comparabilist, I think the universe is essentially deterministic, give or take a quantum stuff, but that doesn't affect free will at all because free will is something we talk about in regards of well should I have done that differently or I could have done that and I should feel bad about not having done it or I might be held responsible for my actions under some circumstances. Similarly personal identity is kind of a macro thing happening when we talk about our lives and interact. But in a world where you can duplicate yourself or merge, people are fluid and in that case personal identity stops being a powerful useful concept. We might need other more subtle concepts instead. That's going to be fun to develop, but it's probably going to be terribly annoying for a long while as we come to terms on how to do it. It's a little bit like politeness with cell-phones in the nineties. It was a long running struggle every society had when it is appropriate to talk on the phone and we eventually end up with various norms and things so now roughly most people know how to behave even if nobody has really explained to them and similarly we're going to have something that makes sense for personal identity. It's just cultural evolution.

FILIPPE _ OK! That's it, I don't want to take anymore of your time and...

ANDERS _ No problem! I probably should respond to that phone call but it's been great fun and I hope you got good material!

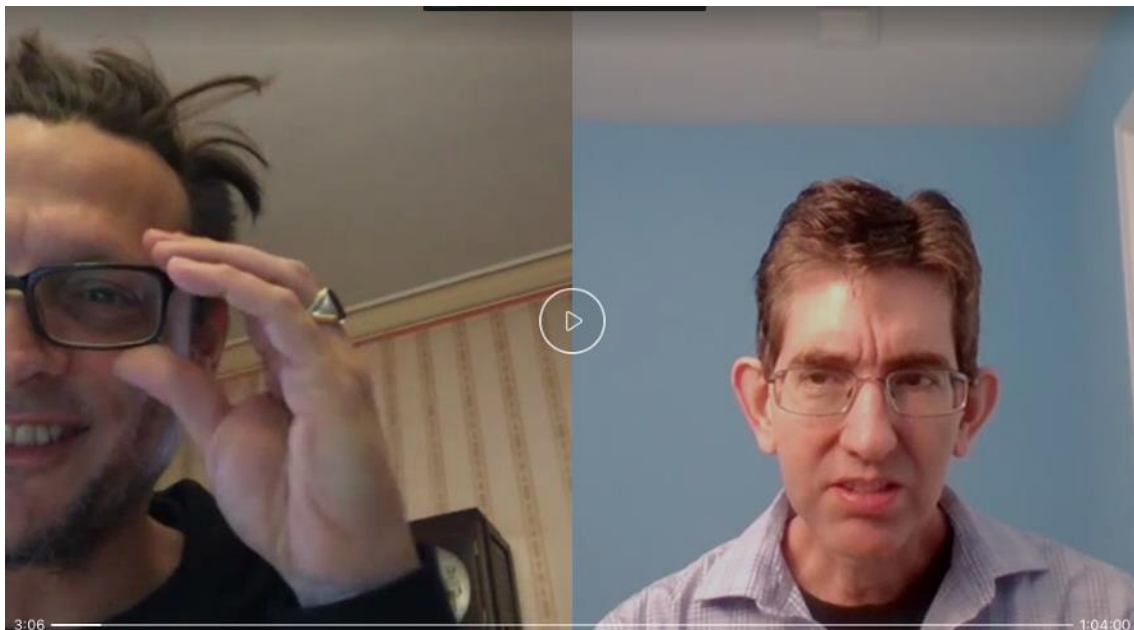
FILIPPE _ Sure! Great answers, thanks a lot, and it was a pleasure..

ANDERS _ Thank you!. OK.

ANDERS _ OK, Bye bye. THE END

APPENDIX G INTERVIEW WITH KENNETH HAYWORTH

(Neuroscientist / Founder of The Brain Preservation Foundation,
USA) [skype@28MAR2020~60min](#)



FILIPE _ Regarding Whole Brain Emulation and mind-uploading. Would another me be me? (or just a copy of myself)

QUESTION I

KEN _ Yes. I think it's a good question and I think that the answer to that question is that there is no me. The idea of a unique soul being part of an individual is a primitive concept that really does not stand up to our current understanding of neuroscience and the brain. So once you accept that and once you're essentially say ok we are what and where does our sense of self come from where does our consciousness come from as, and I would say there are forms of computation of the brain where consciousness comes from and our sense of the brain and our sense of identity comes from episodic memories that we refer to, they're coming from our brain. Once you accept that more scientific definition of the self then you can ask the same question and the answer would be yes a WBE would be you in that sense. Just like if I would copy my hardware from a laptop and put it in another laptop and I asked some relevant question like is this the same laptop, the answer would be well if we're counting about the data then of course it would be the same laptop.

FILIPE _ What would represent the biggest risk or danger in achieving something like MU?

KEN _ Well I think that the biggest risk is that it might work! And you might end up ending in a situation that is bad. That may not answer the question though. Are you saying the biggest risk if it would work or if it wouldn't?

FILIPE _ No no. That if it would work, yes.

KEN _ Yes. I'm very sympathetic with people that lay off the arguments for brain preservation and Mind Uploading, I'll say to my neuroscience colleagues, based upon what you know would you say Whole Brain Emulation would ever happen? Usually they will say yes, if I open up the time-frame as wide as possible they'll say yes. If I say given what we know and our current methods of preservation if sufficient memories eventually they will say yes this seems to cover all the basics. But this would not mean that for society this would be a personal choice and they quickly point out that this could be a

horrific thing and a future society could put in a Zoo... So that's a possibility and there are many many arguments beyond the scientific.

FILIPE _ Moving on to the 2nd question. Regarding Superintelligence and the Technological Singularity. If and when it happens, and we're still humans, would we even acknowledge? Because it could develop so fast in a time-frame of just a few seconds and we wouldn't even be aware of it...

QUESTION II

KEN _ I don't think that would happen. I think we would be aware of it because I see this more as a gradual thing. One of the people that I've been listening to on discussions like this is Gary Markus, who has pointed out (in his book) that the AI in its current state, it had great leaps but there are it completely fails at things that no human being would fail at. Essentially no common sense. And I think that what we will see is more of a gradual progression where we will be extremely aware of, you know this year we got cars that drive by themselves at this level but ten years from now we will get a level where we could actually trust and twenty years later there will be natural language processing where we can essentially ask google what we're looking for and google will understand what we're looking for as opposed of doing a simple word search. So I think it'll be gradual. Although I have to say I haven't given it too much thought. I'll be interested in hearing what you'll take on that but I'll leave it until you're ready.

FILIPE _ Ok still in this question (a sub-questions) how possible would it be for a singleton of some kind to rule the world? Imagine google would develop so fast in that direction that it would make all decisions for us.

KEN _ I have to say I haven't been thinking about these issues (laughs) but I'll give you my opinion. This kind of goes into dystopia. I do worry that the human race has never dealt with really extreme differences in intelligence. We dealt with extreme differences in Power I suppose but not in intelligence. We are a single species and all human beings are all much the same when you're talking about the level we're talking about in the future. So could a singleton (general term?), I think that could happen, yes. I think that, if I can shear a bit?

FILIPE _ Please do you? Sure!

KEN _ The first question you asked me has to do with individuality. I think that the human race is built around this biological notion of a single self and there is no ability for us to really share different memories and to really incorporate different experiences into us and if we're talking about a world where MU has been perfected, then all aspects of the mind will be available for modification. So I could absolutely imagine people sharing memories, changing their individuality theoretically. And a question like 'do you think a single individual taking over the world? It may be that...

FILIPE _ Yes you think that the individual would be all of us in a cloud or something?

KEN _ Exactly I think it would be even more bizarre than a single 'superman' taking over, which is scary...

FILIPE _ Yes the future is scary. Well the present too for that matter (laughs)

KEN _ Yes well the idea behind being scared of this future and that we're going to lock ourselves away from a cure to all diseases including mental diseases. Lock ourselves away from a range of experiences that biological evolution simply hasn't given us access to. Lock us away from space travel etc... Those advantages have to be sacked up against the scary aspects.

QUESTION III

FILIPE _ Can you imagine a posthuman era where we'd be merged with machines and become a kind of cyber sapiens and we'd by then be definitely more bit than bio?

KEN _ Yeah I can definitely imagine that. I think it is inevitable that the human race is going toward MU. I think that's completely inevitable. The time frame is what is in question. I definitely would disagree with people that think that's going to happen relatively quickly.

FILIPE _ Yes I've seen the scale you made where you put it in a not so recent future as opposed to kurzweil's timeframe. He's more "optimistic" in the sense that he would put it in a time frame of about 10 years or so whereas you would go for a few decades...

KEN _ Ok just as a reminder. How I typically think about this is that in the next 50 years I think the human race will get better and better in scanning and understanding brains

and I see a moonshot level program for uploading the first human brain. That would maybe be able to be completed in about 50 years but it might even be a hundred (100). So within this century I think we will upload the first human person. But that could be toward the end of this century in my timeframe and I think it will be a very large project that humanity is going through in several decades and it will not work immediately it will be a painstaking “oh we got her, let’s say her, visual system to work and I’m really happy about that and I wrote 5 papers about how her visual cortex is back online but she’s not conscious of course yet, and now we’re moving on to the hippocampus to see if we can get that back to work and after a couple of decades of that type of research “oh we’re ready to see if we can rebooter consciousness. I’m thinking about that kind of project and then once that is successful in the sense of wow this person has essentially come back, and we think they’re the same person, they say they are conscious then we will, say that they another decade or two of philosophical debate with this MU where we’re really exploring the boundaries of conscious experience because we got the first emulation. Of course this volunteer would have to agree with all of this. And after that process, that again may take the majority of the century, we would really understand how to upload people and even then it would take many more decades to scale it up, to make a routine out of MU in an average human.

FILIPPE _ But don’t you think that, I’m risking an expression here - too linear? Because I could try to convince you with quantum computing that is going to accelerate the means and capacity of computers and AI, and I’d say that, we are today around 90% of the brain emulation of mice, more or less...

KEN _ I would disagree, but I see your point...

FILIPPE _ Yes because one thing leads to the other and unlocks the other and stuff gets combined and works like a catalyst and would boost each other...

KEN _ I agree and for example there are many cases where people would say, so I do connectomics. I do electron microscopy of brain tissue. And I’ve definitely seen projections of people that don’t take into account technology is on the horizon and how that technology will work synergistically to accelerate the whole package. One example of this is tracing neurons. 10 years ago, actually just 5 years ago, tracing of neurons was pretty much a manual process and people would laugh at the idea that we would be able to trace a whole human brain because we spent thousands of man-hours just tracing

hundred by hundreds of microns of piece of tissue and anybody that steps back and says you know once the artificial intelligence algorithms that are being applied to vision, deep learning etc... Once those are really applied in detail to this problem of tracing neurons from volumetric microscopy data that will block people away that people just can't see that that hasn't happened yet. In the last five years that's happened and google has these massive projects for automatically tracing brain tissues. And so there are many things like that and I completely understand that the curve could be much accelerated but I guess Kurzweil, I never actually talked to Kurzweil, but I'm sure that he would say that I'm trapped in that linear thinking.

FILIPE_ I'm sure he would say that yes...

KEN _ (laughs) I'm trapped in that linear thinking. But

FILIPE _ In a way it makes sense that nowadays you have all these technologies working together, I mean 50 Years ago you had disciplines and sciences working apart from each other. Now they are synergistically combined. So now you have nanotechnology, you have biotechnology, you have artificial intelligence, and putting them together and things really accelerate. It makes some sense to me that also because we are a slow natural evolution, biological evolution of humankind if you will, which is quite different from technological progress that in fact is exponential. But we take millions of years to go from here to there, in terms of nature and evolution. Also brains, you know this better than I do of course, have this linear thinking kind of pattern ingrained in our gene pool for billions of years to go from one point to the other. But now things are glueing together and we could shorten that timeframe of yours and come up with brain emulation more quickly.

KEN _ Yes, it also depends on what level of emulation it has to be. And so when I'm debating this with neuroscientists it becomes pretty clear that there is not a consensus on what level would be required for WBE and some of the levels that people suggest no matter how fast we're progressing it would still take a lot of time. I think that people should understand from my perspective. I am way less concerned about timeframes because my main stick if you will is brain preservation. So I look at it and say ok we have the technology today to do brain preservation. And so the argument of whether we should be preserving our brains or not is decoupled from the speed at which the whole brain emulation will come to success.

FILIPPE _ OK. Could you consider that, this kind of cyber sapiens would be the next natural step in evolution?

KEN _ Absolutely. The people that think in contacting alien civilizations. People that really think about it as opposed to

FILIPPE _ Like SETI... (Search for ExtraTerrestrial Intelligence)

KEN _ It seems that they have accepted that if a civilization is way more advanced that it is more than likely to be technological. They have to undergo the transition of whatever their biology is to a new substrate that doesn't have the same limitations. So I really think that it has a logical progression. It doesn't take too much to understand that human beings are not going to willingly accept less than what they could get and that stiking with biology is always going to be less than what you could get by ..designed a new form.

FILIPPE _ Yes because there is also a thin line between medical and therapeutic reasons with those of enhancements. So still in this question you think that politics in a few years could be divided by

KEN _ So it hinges on the few years. (laughs).

FILIPPE _ Exactly because of these enhancements progressing there will be a time when eventually people and governments have to think about who has access to specific enhancements such as MU. Who will have access to MU? The rich?

KEN _ Yeah. So I guess this goes back to my timeline. So realizing I could be wildly off and this is a datapoint, I'm anticipating whatever my opinion is, is going to be on some kind of spectrum of the people you've interviewed.

FILIPPE _ That I've already interviewed? (laughs) I can tell you that I've interviewd Randal Koene (your friend), David Chalmers, Anders Sandberg and now you! Ken Hayworth!

KEN _ Awesome. Ok, I don't know how I'll fit in that spectrum. Let me put it this way. I think the best way to characterize my timeline is I don't think I will see the really exciting stuff in my lifetime. And that is disappointing to me. I should say I've already seen some exciting stuff in my lifetime. That we're imaging pieces of brain tissue at 10 by 10 nm in scale. I feel lucky but at the same time will I see a truly good Brain-Computer Interface? Like what Elon Musk is talking about. Will I see that common sense AI that it can drive a car but also in any kind of situation because it knows what to look for as opposed to .. ?

I would like to think we're progressing very quickly but at the same time the human lifespan is very finite. And I feel like I'm going to miss a lot of the interesting stuff and that includes the question that you asked that politics at some point in the future will be those who want to go faster into Brain - Computer - Interfaces and MU and Intelligence augmentation (or whatever you want to call it), artificial intelligence in terms of incorporated data to our daily lives (general intelligence I mean), there will be people that want to do that and people that have seen all the movies that tell you that's a bad thing to do (laughs) and try to slow that down. So right now if a politician said we need to really worry about AGI maybe taking over the world in a few decades everybody is going to laugh because we don't have artificial general intelligence and that's not really a debate for right now but will be probably sooner than most people would probably think but not right now. Later in the century we will have that.

FILIPPE _ You're sticking with your linear thinking.

KEN _ I'm sticking with *the linear thinking* (laughs).

QUESTION IV

FILIPPE _ Regarding Superintelligence and Mind Uploading, I have this kind of thought experiment. Would it make sense to iterate human intelligence augment and improve it and in this manner cope with AI's progress?

KEN _ What do you mean by iterate?

FILIPPE _ OK. Let's take for instance a SIM1 from me, I would make a copy...

KEN _ Ok. You would make many copies....

FILIPPE _ Yes, and I could edit them, like you said in the beginning of this talk you said that there's no actually a real 'self' so regarding mind uploading I would still be me. So I would make a hundred copies or a million, but I could alter them, change them, by editing them, just like we do in genetic sequencing nowadays, and I could then confront these various copies in a simulation with different scenarios. Economic, Environmental whichever, for instance I could in this manner, make a graph, a plot, and from the responses I would get from my different uploads choose (select) the most adapted one, the most intelligent one. By playing chess or go if you will, and so find the most eligible.

KEN _ Ok. I think that people will definitely do that, yes. But I think the key question here is what level of consciousness is required for general intelligence, ok? The human beings brains, and when I say human beings I mean the mind uploaded as well. So our conception of a MU is essentially a human being but emulated in a computer. So that package of the human mind is more useful than any artificial intelligence... (clockwall)

KEN _ (no problem) So our idea of AI today is way lower than human beings because we have this general intelligence (common sense and stuff like that etc...). So when you're saying to iterate that, that could be a very very powerful augmentation because now you have a thousand "yous", momentarily, working at the same time for the same problem potentially from different aspects. This of course assuming that whole brain emulation can be made relatively inexpensively, which is a big assumption, but...

FILIPPE _ By that time it might be quite cheaper than it is today...

KEN _ Exactly at some point this will be the case. So the reason why I recoil from that is because I think that, I'm not sure about this, this is a question that I would love to have somebody answer, maybe someday already answered this, but how much of our common sense and general intelligence is tied to our consciousness? Can we have the type of really subtle common sense intelligence without bringing things like suffering and the sense of self and things like that. And if I had to guess, I suspect that those are decoupled, but I'm not sure at all. And so iterating under that scenario where you would decouple your general intelligence, you would never want to iterate yourself because it's inefficient. It's like saying I created this algorithm to do a particular task but I'm having to do this visual display that has never been seen by anyone. And under those circumstances why would you do that. So, that's my subtle answer to that question.

FILIPPE _ OK. But we can develop superintelligence, or even just an intelligence for that matter, without developing consciousness.

KEN _ Yes.

FILIPPE _ Those are different issues. Assuming consciousness would be a pattern, an organizational invariant, it then wouldn't be substrate dependent, whether biological or silicon. In fact in this view a computational simulation would be able to emulate that same kind of pattern. So consciousness could be achieved in this manner. Nevertheless, I

know what you mean, all those millions of copies would be conscious, but we could iterate our intelligence this way and make us more adapted.

KEN _ Absolutely. And I don't think that we would have any problem re-emerging those intelligences together.

FILIPPE _ Right, if we can upload them then we could also download them, I imagine. But this could also lead to something else, some kind of a predictive response upon specific scenarios, such as war, or like now viral pandemic scenarios, so if we could iterate these intelligences we could also predict our own responses to these kinds of scenarios and in this manner maybe avoid disasters.

KEN _ Yes. Let me ask you a question about this. I think I'm on your wave..on that, ok? To me there's two types of people. Or maybe a spectrum. There's a type of people that think that WBE is going to require a simulation at the ion-channel and receptor, molecular level and somehow the human brain is using the conformations of individual proteins are absolutely fundamental to the algorithms which of course they're absolutely fundamental from an implementation level, nobody would doubt that. But from a computational level there are people that think that there really is a crucial computational level, and then there are people like me that have come from more of a ?? science background, I was deep into the symbolic AI literature and computational models of the brain that abstract many many things away. So long story short there are people that would think that a human brain boiling down its algorithms, its algorithmic complexity might come up with something that could be a regular sized supercomputer today versus people that would think that it would take billions of supercomputers mapping out the conformations of individual proteins. And when you pose a question like this to those different people you're going to get wildly different answers even if maybe that is irrelevant. You're still going to get wildly different answers. Because when you say iteration to somebody that thinks along the 'oh this is just pretty algorithmic simple to begin with, they're not scared of the idea of iteration. But when you think about somebody that says that we are going to emulate every single molecule, the idea of iteration is laughable because they think you're hardly even going to get to the one emulation. So I'm very much in the camp? that algorithmic complexity is not very large to begin with and therefore I think iteration is something that is going to be doable sooner rather than later. Sooner might be in a hundred years. (laughs). Let me put it this way I think we could emulate a brain really

well if we'd understand how it works. Our computers are capable of the algorithmic complexity. We just can't program them.

FILIPPE _ Yes, but actually we could emulate it even without really understanding it. It's logically possible that we achieve the time where we make stuff and do stuff without really understanding them. I mean I'm convinced that we don't have to fully understand the brain to actually emulate it. I'm not saying it's good or bad, just think that's the way things are.

KEN _ I agree with that argument, it's true but I think that we will understand the brain before we get to the technology to emulate it. I think we will reach a completely satisfying algorithmic understanding of the brain. We wouldn't have to. But I think we will. Because I ... in about my field.

FILIPPE _ I would say it's safer to first fully understand it before emulating it. Still in this question with the thought experiment, I guess you already answered this next sub-question right at the beginning because imagine this scenario of copying millions of me's, and I have to believe (I don't like the word believe very much but) that the copy is 100% equal to the original and let's imagine it's a blueprint just like the genetic analysis today, and then you can compare these genetic blueprints from different persons, for sake of the argument my and your genetic sequences would be 99% similar and now imagine this in our MU scenario that we would compare our brain-blueprints and now imagine the mindblowing case that our blueprints would match 100%, so they'd be 100% identical. Would this mean we're the same person? (laughs). I know we're not but would make my case more difficult in claiming we are different if we'd share the same MU blueprint...

This creates quite an issue in terms of personal identity don't you think? But as you said at the beginning of this talk of ours, it doesn't really represent an issue to you in terms of your standpoint regarding personal identity...

KEN _ That's right. I think this is mind boggling to think about. And even though, I could come in and say - look, every single one of these questions that involves personal identity and mind uploading and the philosophical copy problem. If you would ask that same question to say C3PO, some robot with a computer brain, in that setting when you talk about C3POs computer brain copy happens to be identical to somebody etc etc ... If there's a simple answer from that second person, or third person view of a robot, and yet there's some demystery when it comes to human beings you would reach rocket science,

essentially. Because we are just biological machines. So whatever works in the C3PO example should work for us as well. At one level that answers all the questions. But on another level it raises a whole bunch of new questions. Ok now that you accept that what is the .. So let me rephrase that is not hypothetical at all. We have two inventors. And one of them invents something really fantastic and somebody else really no connection, invents exactly the same thing. So who's invention is it? Is it the same invention? There's no answer to that you know, but the same...

FILIPPE _ Yes, it happens all the time...

KEN _ Yes and I think the same thing would happen here. Now I think this idea that we have 100% compatibility. I think that's never going to happen. Because I could say, I wrote this book, and I sent it to the publisher. It's a 500 hundred page novel. And at the same time somebody sends to the publisher exactly the same book. Word for word. Character for character. What is the probability of that? The answer is it's never going to happen. That is ludicrous. That is never going to happen. And that's the same thing when it comes to a human being. Maybe I think you're asking a more subtle question. If you could get the case where you would get the same book in essence to the publisher. Yes the words are different. Yes the main characters are named differently. But the subtle plot twist that really matters is the character development that really matters to the story and not the detailed words they're using, those are really the same because that's less information to the side. .. And I think that at some level, because I think that there are 7 billion people on the Planet right now, you could argue that there's a KEN rolling around there that is not identical to me that has exactly the same memories as me but from an essential essence point of view it's good enough. Why should I care about my existence as long as number two is running around. And there's no good answer to that. I think that is not a scientific question, it is a human experience type of question. It still deserves serious consideration.

FILIPPE _ Yes, because people like to think they're unique. Until some extent this could end up with that 'uniqueness', in such a mindblowing case as this. Ok last question and we already covered this over our talk. Regarding transcending biology, is it inevitably to generate a dystopian society?

QUESTION V

KEN _ No, I'm definitely going to say no! (laughs). I mean, it could generate a dystopian society.

FILIPE _ It also depends on what we consider dystopian right? ... We're now already living in a kind of dystopian world anyway.

KEN _ Right. I'm definitely along the lines, I mean, personally I don't know. Again, I'm not making some kind of claim about the logical progression of societies through a technological singularity or anything like that, I'm not saying that.

FILIPE _ Yes, well I guess there's a subtle sub-question inside here that is why I address this dystopian society issue because in terms of what enhancements will bring us in the next years could lead to those people who would live with such enhancements and upgraded that they'd be way beyond those that simply don't have any kind of upgrades, you know? And it could make such a huge difference, a disruption in the fabric of humankind. You could end up having cyber sapiens walking around with homo sapiens. And by the end, if darwinism is right, probably cyber sapiens would make it and we would be outcast.

KEN _ That's right. I think that, as you were saying, things are always connected. So what we're imagining, and this is going to be kind of outlandish. What we're imagining is a world where everybody truly understands how the brain works. The society at least understands how the brain works. And at a very deep level. Not today's society. Where most people think they have immortal souls that are dropped from heaven or something and they really don't understand this subtlety of uniqueness. In that world there is a different philosophy of morality. We haven't seen it yet. But I think it will come about. In that world. So for example if I make a copy of myself, as MU, and I see that that copy is being tortured, how should my feelings about that happening be? Because my point of view is not feeling that torture. And I could potentially say, look I'm not going to reintegrate those memories. Screw that! I don't want to have those memories, ok? So I got a copy of me that is being tortured and I'm not going to remember those should I have no quarrel about that happening? I hope the idea is: No that would be horrific, right? I would go out of my way just as if somebody was torturing me, I would go out of my way to stop that torture from happening because in a certain sense in almost every sense that is me being tortured, right? Now move that just a little bit beyond. Let's say it's not a perfect copy of

me. But it's a reasonably good copy of me, it's a copy of me from three years ago, it had some degradation, it had some new life experiences but it's still being tortured. What should my relationship be to that? I would hope, still think, that's me in essence. So what I'm saying is that...

FILIPPE _ I would definitely have that decision, like most of us, I even would do this with my parents, or sister, and they're not me. Close, but they're not, and I would still do it.

KEN _ So I guess what I'm saying is that one of the perhaps logical conclusions of our really truly understanding the nature that we are, that we are computational, and our self-identity is really tied to a set of memories and not a point of view is to accept the idea that we're all the same person. And that's a philosophy that's been around and I've written about it a little bit. But it's a philosophy that I think will gain more and more traction as we become more comfortable with our scientific understanding of who we are and therefore I think society will become less and less dystopian. Because we will be adopting not a philosophy based on nothing but a philosophy based on science that we are all the same individual. In a way that's not saying that we should all be one, you know that there should only be just one of us, or that we are all in psychic communication with each other, it's not saying anything like that, but it is saying that just as we would abject to our future self being mistreated, we should abject to somebody that we have never met, being mistreated and that there is no sense in us trading of a personal gain versus the suffering of somebody else because we're essentially causing ourselves to some extent to suffer. And I'd like to think that the dystopian possibility will become more and more remote because of that type of thing. I could be completely wrong!

FILIPPE _ You could. (laughs)

KEN _ (laughs) I could, definitely.

FILIPPE _ OK. Ken, we could go on the whole afternoon, like you said in the email such themes are always so exciting and interesting...

KEN _ Yes. It sounds like a fascinating PhD project! I hope you'll send it to me when you're done.

FILIPPE _ Sure! Yes. I will keep you updated. So thanks a lot and I'll be seeing you!

KEN _ Ok. Yes! Anytime. Thank you! **THE END**

skype@28MAR2020~60min



FILIPPE _ So let me start by thanking you. Regarding WBE and MU. Would another me be me? Or just a copy of myself?

QUESTION 1

NICK _ Wow. You realise I'm an academic philosopher. What are the odds when you get a question like that? (laughs). Trying ways not to answer it. So, would another me be me? I

Guess so" I think that, in my view, there's a difference between copying and a set of processes that suffice to transmit identity or survival. So I am convinced by this ..? that says you could make a copy of you, and you could copy me precisely, but I shouldn't view that copy as a continuer of me or as a, I mean obviously a copy of me would share many of my interests and I'm sure I might quite like that copy (or probably I wouldn't like that copy), but it wouldn't be me. One of my papers was in a OAP? Volume. I mean suppose you were to learn, I mean philosophers love to write thought experiments, so please excuse me. So suppose you were to learn that there are extraterrestrials with powerful scanning technologies and they've been scanning Earth from a great distance for quite some time and also they've managed to basically (they've become very interested in you!) and they have basically produced many replicas of you. I don't know what they do with these replicas. So, they treat them quite badly, because of their experiments. It seems to me you might feel sympathy with these replicas if they have been treated badly. But I don't think, I mean if you were told wow these replicas of you, exact copies have been tortured I don't think you should complain, I mean you might complain about these extraterrestrials but you shouldn't say oh that was me.

FILIPPE _ Yes, but probably those replicas would say exactly the same and say no, I'm me...

NICK _ I do think that's similar to cases, I mean there are people who believe they are Napoleon. Who knows? There are some great books about Napoleon and they may have done some intense study in all of that. It seems to me that no one today, no matter how much they would replicate Napoleon Bonaparte's psychology is Napoleon Bonaparte. So, I guess I disagree with views like the psychological continuity type of personal identity.

FILIPE _ OK. What would you think would represent the biggest risk/danger in achieving something like this? (MU) (risk for society in general).

NICK _ I would think again. I mean it's a general kind of concern. I mean I'm not saying that I don't know, if you would upload me, into a digital copy, I don't see any reason why it couldn't be conscious. And why it couldn't go on to live a wonderful Life. But again it's kind of that issue of well I'm still here, and then there's a digital copy of me that maybe starts out sharing my interests and wanting to preserve me, but over time develops quite good interests and notices and says sort of well, this earlier human being from whom I was copied is quite susceptible to this coronaviruses, I'm not susceptible, and do I really care? Well a little bit, but not much. So, I guess as soon as you do this you do almost inevitably create a being that has certain interests. And I don't know if I would want to. So, my children for example. I guess I care more about their interests (they are flesh and blood), than I would do right now about a digital me. And I wouldn't want a digital me to say well screw your dumn children!

FILIPE _ Yes. But I also mean for Society, you know, the risks and danger in achieving something like this with a millionfold hypothesis that it could unchain within Society...

NICK _ Yes it could be great, you know, if anything it could be great. I guess you can be in danger as in too much optimism. So there's a danger in saying, I mean the very idea of facebook, could be great. Could bring us all closer together and not empower terrorists. You can still say, I've finished a great book, by Steven Levy, a sort of counter facebook. Sort of talking about this still idealistic Mark Zuckerberg. And it's sort of in a way yes!! mean he's still hanging on to this idea that facebook will make us one big happy group hugging the world. And it's nice to have that. But I suppose when you have a technological ideal, I think it's incumbent on you, it's important that you would ask what happens if we would fall short? Any technology is worth pursuing, but I guess you should be aware, I don't know of near misses? And that achieving that ideal can be pretty disastrous. So, I would ask the same general questions about uploading as soon as we (if we) would succeed. Then you'd have a lot of possibly considerable morally beings who say well, you flesh and blood types, I mean I actually don't believe that I am identical to...

FILIPE _ And with rights. They would have rights.

NICK _ Yes. So, I can't just destroy them. That would be wrong.

QUESTION II

FILIPE _ Ok. 2nd question and regarding superintelligence and an eventual technological singularity. If and when it happens, and we're still humans, would we even be aware of it? Imagine AGI would develop so fast in a time frame of about a few seconds we wouldn't even be aware of it. Would we understand it? We'd be like bacteria on Mars in just a few seconds, you know what I mean?

NICK _ Yes. Or maybe more slatternly we would be the australopithecus.

FILIPE _ Yes...

NICK _ Well, Nick Bostrom's book on superintelligence, it seems to me that if I had a complaint about it, it would be that it has too many arguments in it. I mean if you've seen the terminator movies and skynet, then you have seen enough to say oh, we got to watch out for this. So, he tries to argue in that book that this humanity destroying AI or an artificial superintelligence is kind of inevitable or it's a devolve. A kind of thing that will happen! And my credencies by degrees of belief, well I'm more scared of climate change from an immediate extinction than, well I don't know how it will happen, but than I am from this laptop that I'm speaking to giving rise to something like an AGI and then decides I don't need you anymore. I'm not a specialist on that, but I'm not particularly scared of ET turning up and doing this "Independence Day" and wipe us out either. I'm sure it could happen...

FILIPE _ Still in this question imagine it would develop into such an extent to become a singleton empowered to rule the world. And we'd willingly give our decisions, you know to a better decision maker. We already do this today with google. I know I have to turn left here but if google's saying no, I'm going to stick with google's decision.

NICK _ I can envisage this. When Bostrom talks about, was it the Oracle?, so it's sort of interesting how, I mean it's uncandy isn't it? I have a bad sense of direction and when I travel to foreign places I don't even bother to trust my nose because I know that my instincts of which way we should walk is the very opposite. So, I always trust google over me. So, I think that is quite likely. And I think... I don't know how long it will take before that kind of AI can write a better philosophy paper than me. If you think about all the tasks in writing philosophy, being aware of all this literature and knowing the points, I'm sure it's sooner than we think.

FILIPE _ For sure. Just a few years ago, two or three there was a contest in writing sci-fi short stories in Japan, the winner actually was an AI, and ended with the AI claiming some will (this seems too sci-fi now, also because we're not talking of AGI yet. It's not that we've reached a *conscious* AI or something yet, but nevertheless they're also painting stuff now, and ending Beethoven's 10th symphony, so there are reasons to believe that AI could in fact more sooner than later become such a thing. But the question was actually if it would be possible to let things slip our hands in such a manner that something like a sort of singleton of some kind would in fact be able to rule us.

NICK _ I mean it is possible sort of in a way that anything's possible. But I wrote just a paper on this. So, here's why I don't feel especially worried about that. And the paper's called "Don't worry about superintelligence" and basically the line is yes, it is possible. And there are lots of other things that are possible. Every time I cross the road, a violent death could hit me, and because I'm not perfect and I'm eligible to take cars so I'm academic and choosing to cross the road exposes me to some risk. And it seems to me that people like Nick Bostrom and Anders (Sandberg), who is a very nice fellow by the way, make too much out of it. So, in a way if I was to distribute my worry, then worrying about crossing the road and even if I have a non-zero probability of getting killed, somehow it can be right to take that risk, if you think there's some nice food over there. So I guess I would think the same so how worried am I about the singleton that might say that I got rid of the other AI, like google because that was turning into something, and now the humans they made google, they're now my potential threat... So, you know that spooky sort of sequence that scene in Terminator where it describes skynet and it kind of abruptly becomes conscious and 5 seconds later launches the nukes...

FILIPE _ Yes, I think it's the 2nd Terminator.

NICK _ It is the second. It's a great one to sort of sight... In terms of what AIs can do it's a lot easier for Google to answer. I bet google could write me a great paper on Kant, maybe not now but soon, some google successor if it wants to. I mean it's going to take a while just for moving around the environment. I was impressed by the Morovitz paradox, which you might know quite well. The hard things are easy and the easy things are hard. I mean the AI that controls our missiles and wants to form a plan to act on the world to achieve this end, that's quite difficult. So, I don't feel that scared of google one minute being a weather forecaster and the next it's launching nukes. I don't feel that

scared. So, when I set my credencies I don't feel so scared, I mean it's possible but I guess a supernova forming in our sun is possible in some sense...

FILIPPE _ No, actually not. Not our sun.

NICK _ Given what we know now. We could be mistaken.

FILIPPE _ It's a too little star for that. It's just an average low-medium sized star.

NICK _ Yes, I know it has to be big and it turns into a red giant...

FILIPPE _ No I don't think so, it's also not turning into a red giant. But I know what you mean, yes.

NICK _ We could be wrong. That's what the current model says.... I read a book just a while ago over a viral spillover and bottom line it was talking about SARS and I always think that it's important for us to be aware of the current model. I know our sun is too small to become anything like that and if it did, we'd have plenty of warning. But that's according to Albert's? theories. They could be mistaken. So, I know there's like zero probability now in the sun exploding...

FILIPPE _ Yes, there are so many different kinds of measurements made from different areas. You got them from physics, but also from chemistry and so on and soon... This is actually one of the good things about science or that I like, is that you can try to prove or even prove something in different ways. Like I'm pretty much sure that the Earth is about 4.6 billion years old because we can prove this chemically, physically, biologically in so many different ways. I'm not convinced that the bible's right when it says that the Earth is about 4 thousand years old.

NICK _ Yes when I was talking about credencies, that is the biblical story of the solar system .. it has a nonzero but very very low probability of being true. The world would have to be crazy. There was this great documentary on Netflix on flat earthers. People who literally believe that the Earth is flat. It's an insane view, but I guess given the evidence that we have. Nonzero.

FILIPPE _ That's "America". Not to stereotype americans in any way but they do have these kinds of things. It's a movement actually going on there that thinks the Earth is flat. And there are lots of people over there that truly believe the Earth is flat.

NICK _ It's tragic isn't it? And when you see these things on netflix, and these people seem quite sincere. They're not just idiots but ... And

FILIPPE _ You know that they made a census in America asking people what is more important. The sun or the moon. The census was quite big, and the majority said the moon. Because the sun comes up when there's light anyway and the moon gives you light by night.

NICK _ (laughs). It just shows that there are certain questions you shouldn't ask people.

FILIPPE _ Well, not in America at least. Do you think politics will be divided between bioconservatives and transhumanists in the future? (clockwall rings)

QUESTION III

NICK _ That was impressive (the clock banging). That seemed to underline that question (laughs). That is an interesting question. So everytime anybody tries to predict politics I'm always reminded of Francis Fukuyama. You know Fukuyama "The end of time"?"

FILIPPE _ Yes. Sure!

NICK _ I remember that book and it all seemed quite plausible to me at the time. So, I'm always quite weary whenever they're predicting politics. I do think that when it comes to the bioconservative rejection of human enhancement. But that ship already sailed. It doesn't even make sense to sort of say oh I'm going to reject all the enhancements, because many of them are just part of our ordinary existence. You know the old comparison of genetic enhancements once we achieve this how different is that from educational enhancement. Enhancement is part of our culture. The interesting thing that it makes it sort of difficult to work out which ones we think we want and which ones we think we don't. I would like that to be a collective decision. Where J.F. Bejos and his billionaire mates say well it suits us to allow ourselves to have those options. But that's just for us.

FILIPPE _ Yes but what's really in this question, too, is that if we can imagine a kind of a cyber sapiens could become with all these enhancements in some years and we'd almost be more bit than bio, and somewhere in time I don't know ten, twenty years whatever, you could end up in a kind of dystopian society where you'd have people upgraded at

the same time, and those who don't. Because who's going to empower these enhancements and also who's going to govern those enhancements? So classical left and right wing wouldn't be there anymore but instead bioconservatives and transhumanists because of this kind of cleavage in society?

NICK _ I think that's a real possibility, yes. And as these technologies become more available and more powerful because I guess in a way there's always been this sort of argument that sort of says we already, I think Alan Buchanan has this term maybe he's on the best use of risk pioneers. So when you look at these exciting enhancements like oh I could attach a chip to my prefrontal cortex and see what happens what can I do, that there are people who want to take that risk. And it's kind of good for us, so that then they can say well I used my own cash to this, I'm my own risk pioneer. I'm like the Wright Brothers. I'm going to take a personal risk and if it turns out badly, well that's bad for me, but I have taken the risk. So I guess that would be good, I mean the idea is, according to Buchanan, then once that works then a technological diffusion happens. But I think it's naive to assume that that's, I don't know, that very expensive cancer drive that you got now, that only the rich can afford inevitably gives rise in a few years time to a cheaper version, that's more broadly available, and you say thank you for testing this drive for leukemia, now my kids can afford it. Thank you. It's naive to assume that that will necessarily happen that way. So yes, I think it is possible as these technologies become more powerful there's a good chance that the social divisions will be defined around who has access to them and who doesn't.

QUESTION IVa

FILIPE _ Ok. fourth question, and almost the last one. Again regarding superintelligence and mind uploading. This is a little bit of a work in progress of a thought experiment of mine. Would it make sense to iterate human intelligence via MU to augment it to improve it and in this way cope with AI's progress. So what I mean with this. Imagine the scenario you would make an upload, it would be feasible, we could in fact do it. And so I would make copy nr1, of myself, a blueprint of my cognition. And I'm going to make an analogy with nowadays, I particularly work in molecular biology in DNA sequencing, so today you can make a complete sequence of all your DNA and make a comparison with exactly my

DNA. Compare both DNA's, of course there are going to be differences, but in the 'package' both of our blueprints would be 99% similar. But they can't be 100%, of course..

NICK _ No, even those 1% or whatever...

FILIPPE _ Yes, makes a difference.

NICK _ Exactly.

FILIPPE _ Ok, so still in this scenario, I would now make vi MU exactly the same, I would make this blueprint of my brain, and make copy nr1, and since it's digital I can edit it (just like I do today with genes), I can alter it, and change it, so I would change, and I can make all the copies I want to in just a few seconds like any kind of file

NICK _ Ok. Yes, digital technology. It's one of the big selling points.

FILIPPE _ Ok, so now I take copy nr1, Filipe 1 until Filipe 1million. And between them I'm going to alter and edit them, and then I'm going to run all these copies in a software, in a program just like a game, I'm going to expose my copies to different scenarios. They can be environmental scenarios, or just like today we're living a pandemic viral scenario, economic scenarios, whichever. So what I'm then going to do is evaluate the response of those copies of mine, one of those scenarios could be playing chess. So I'm going to make a plot and compare them and I would then take out, select, the best response. So basically I could iterate the most intelligent me, you know what I mean? So, actually , does this make sense to you? Do you think it would make sense to iterate my copies in order to find my best me in playing chess or something?

NICK _ You know what I like about that thought experiment? Is that: It's a great thought experiment! I mean, you described, you said enough about it to make clear that it does not violate any law of physics, I mean it's potentially doable and I guess if you're envisaging sort of, you know it's kind of like, I think that one of the main dangers when we approach these things, it's a bit like with the coronavirus, isn't it? We find ourselves in this situation and wonder wow how do we get here? This is completely novel. So when you have a thought experiment like that, that sort of says , well I mean this is one of the defining features of the digital! The digital is copyable. And it's designed to be precisely modified in a way that in principle we can modify ourselves. We haven't cracked that yet, but we know how to make digital copies. And we can presumably make the changes you

described I guess in a way we try to work out the best digital version of you at chess is the way to do it. I mean ..

FILIPPE _ Yes, well of course the idea here isn't to get the best Kasparov on Karpov, or even deep blue for that matter. It's just that by iterating intelligence you could select, there's also a little bit of darwinism here too, because you're selecting the best response (the more adapted one) in this algorithm.

NICK _ Well, it's a bit like alpha-Go isn't? I mean in a way that essentially you're describing a process that is not too dissimilar from the way, you know the defeat of Li Sa Dol (the korean Go player). Basically just played itself. I guess that the digital copies of you do engage in some sort of interaction in which you're able to sort of say wow this or these 2 copies have produced the result that I value.

FILIPPE _ Exactly, and exposed to different scenarios, viral, environmental scenarios, you could choose the one that responds better, so the next time you get a scenario like the coronavirus you would know how to better respond because you already predicted it.

NICK _ Yes. It's a great thought experiment, isn't it?! I mean if you think it in a way there's a lot of versions of me and let's put them in a situation like climate change so that they can interact with each other and interact with intelligences that are copies of you and that are capable of engage this sort of, I don't know making decisions that will promote welfare and are also not going to cause so much carbon into the air. I mean imagine what that could have a potential approach to these crises.

QUESTION IVb

FILIPPE _ Exactly yes. Ok, so still, and just to finish, in this thought experiment, imagine the mindblowing case, you would now compare my blueprint, upload my copy with yours and imagine the case they'd be 100% identical. Could this mean, wow where's my personal identity? You know? I'm me! I don't want your blueprint to be identical to mine! Because I want mine to be different because it feels so me, and I give such value in being me... So we were to actually achieve MU one day, say 50 years it doesn't matter...

NICK _ Well in the first part of your conjecture, you talked about sort of a quasi darwinian sort of, so you can imagine, Filipe your versions of you, you twig some, I have some

versions of me. I guess they start out just maybe similar in many aspects but then we impose some kind of darwinian process on them, so I suppose that we'd go for the same weightings. When we oppose the same values. Solutions to future corona viruses we want to solve climate problems, so we make sort of modifications. Maybe some of them are random, to make it even more darwinian, but the convergence you talked about, I can imagine that could happen. So in a way a hundred thousand nick agar copies are all slightly different and ...

FILIPPE _ Yes but don't you think this would mess with our own psychological and philosophical values we have regarding personal identity?

NICK _ I think that right now I am flesh and blood. I have a conscious identity that I care about and I would say if you're going to make a copy of me and.. I would say I don't care what you do with that, make the copy and then delete it. It's not me. Please don't do that to me. But in this future it's almost like we'll need new concepts of identity! The digital intelligences will want things and value things. But I imagine I care about a concept of identity that is within my physical self, but I bet they're not going to! And they're going to say well what's the concept of identity that is relevant to the viruses? To the corona viruses? Well obviously they can't conceive of that. But if they could, then it wouldn't be any kind of concept of identity we use for ourselves. So I suspect that would be the same. I mean whatever concepts that we organize our values around won't be relevant or only be partially relevant to the concepts of those digital selves world.

FILIPPE _ OK. That was it! These were my questions. So thanks again and how you're dealing with this coronavirus situation altogether? Do you think it's getting ever worse and the world is going to end? This is

NICK _ So we're all here at home... Well it's sort of funny isn't it? I mean it's not funny. But so I'm teaching this course called "How to study the future". And I kind of presented and it sort of started confident about the future. We're very future blind and future over confident. But I mean who could have predicted this? Every day I wake up and ...

FILIPPE _ You mean you could have predicted this?

NICK _ No. I mean people say we could have predicted it.

FILIPPE _ I read a paper in December just before the outbreak. A Japanese study with a 99% similar virus to the coronavirus. And they made a simulation and at the end 65 million people would have died.

NICK _ Yes it's terrible. Have you ever read a book by David Karmen (he's a popular science writer), called "spill over". And it was published I think 2013. It talked about coronaviruses mainly in reference to SARS and basically said how lucky we were. I remember reading it and thinking wow that would be terrible. Imagine it wasn't ..of spreading and there's nothing to prevent. When anybody talks about these things you suddenly go wow and then of course in a way you can't really predict it.

FILIPPE _ Yes it feels like a sci-fi movie or a *Black Mirror* episode...

NICK _ So I'm trying to teach my students to watch lots of sci-fi and I don't even think sci-fi predicts the future. But I do think it contains lots of vivid descriptions of the future. If someone would have sat down and said what if SARS could spread like the common cold? If people would have sat down and said let's think this through because there's nothing that would have made it impossible for SARS to spread like a cold. And it would have been good if we'd said ok let's think about that. What would we do? Because I think people just said that's too horrible. The death rate of SARS was about 10% or something? Or lower?

FILIPPE _ Yes, lower. But around that yes but less contagious.

NICK _ Yes not that contagious.

FILIPPE _ Well the whole thing got out of proportion because it never was in control actually. Right at the beginning it started bad and currently it's even worse and I think it's going to be a disaster in the US even worse than Europe. Because they don't have a real health care system.

NICK _ Yes and they have an idiotic President that looks terrible.

FILIPPE _ Yes but idiocracy is also the future.

NICK _ (laughs) Yes well let's hope. So there are models of the coronavirus, maybe this isn't good for you, but there are models of the coronavirus in New Zealand that suggest that if we're lucky we might get away with just twenty deaths.

FILIPPE _ Twenty? (20). And if you're not lucky?

NICK _ If we're not we may turn into, well our prime minister said, when things started getting bad in Italy, she formulated some policies well let's try to avoid what's happening in Italy.

FILIPPE _ Right. But you're controlling the situation in New Zealand right?

NICK _ We're trying to but numbers are going up...

FILIPPE _ But the All Blacks aren't playing at all right?

NICK _ (laughs) No no

FILIPPE _ Just another year not winning again is it?

NICK _ (laughs) Just another year not winning again, but I guess the same applies to Bayern Munich or Porto....

FILIPPE _ (laughs) I don't like Bayern nor Porto. ok , Nick this was a great talk we could go on and on but I don't want to take anymore of your time with it and I have my interview done...

NICK _ Thank you! And good luck with the thesis. When do you expect to submit? I hope the coronavirus doesn't stop that!

FILIPPE _ This year for sure. I guess in Fall or so.

NICK _ OK. Maybe if there are international conferences again

FILIPPE _ Yes well actually there was a conference I could have been in Italy a few months ago for a conference but I'm guessing I'm better off this way.

NICK _ Well in a couple of weeks I was supposed to be going to a conference in Belgrade, so that obviously won't happen. Well they said they still want it to happen maybe later this year but I doubt that.

FILIPPE _ No. I doubt that, too. There's not going to be any kind of traveling anymore until I don't know. Ok.

NICK _ Well good luck. Where are you in Germany now?

FILIPPE _ I'm in a little town near Stuttgart. It's my mother's birthplace actually. I'm not from here, I was born in Hamburg, then made high school in Lisbon, and University, too. And then I came here again, so back and forth.. But yes. A little town near Stuttgart, and also at lockdown of course, very quiet and peaceful, I don't see anything around me other than birds and trees actually...

NICK _ Yes we can go out a bit but don't see much in Wellington, but, Ok so good luck with that and I hope Stuttgart stays a good place to be off the coronavirus and I enjoyed talking to you so hopefully I'll see you at a conference sometime.

FILIPPE _ OKok. Me too. OK Bye.

NICK _ Bye bye. **THE END**

—

skype@31MAR2020 ~60min



TOM _ Filipe? Can you hear me?

FILIPPE _ Yes, sure!

TOM _ Salute

FILIPPE _ Hi. I guess my bad. I mixed the time. You said 2pm my time, but actually I'm in Germany so..

TOM _ AH! Wie geht's denn heute?

FILIPPE _ Sorry? Oh. Ach so, Sie können Deutsch!

TOM _ Ja! Ich bin ein Koch!

FILIPPE _ Ja genau! Yeah...

TOM _ I also know some spanish and some "português"

FILIPPE _ (laughs) Oh pá!

TOM _ But we'll do english that will be easier. (I'm just going to turn off the radio).

FILIPPE _ Yes of course. Sure!

TOM _ I'm back. As you may imagine the corona virus has kept me very busy. A lot of my work is in public health. Epidemiology.

FILIPPE _ OK.ok.

TOM _ So what can I do for you?

FILIPPE _ Yes. First of all let me thank you once again for being so friendly and available to me...

TOM _ I'm delighted to help other's work. Especially when they're working on thesis and dissertations.

FILIPPE _ OK.Great. Secondly I have to ask your permission in recording this.

TOM _ You have permission to record! If I'm helpful and we work together at some point I may ask you to send a note to the chairman of my department. Just letting him know I've been assistant and helpful to you in your work at this level of a thesis. He likes to

know I actually do some work. Ok you have my permission and ask me about anything but my elderly sex life!

FILIPPE _ Oh God!

TOM _ (laughs)

FILIPPE _ I thought that was Sci-Fi?!

TOM _ (laughs). I think that way sometimes!

FILIPPE _ I've watched this once in an episode of Family Ties. You remember Archie Bunker?

TOM _ Yes of course.

FILIPPE _ He said that beautiful sentence once regarding a book about sexuality in the later age, I don't know. And so he described it as being science fiction. It's always good to make a quote by Archie Bunker, don't you think?

TOM _ Yes. It wasn't Family Ties. It was All in the Family.

FILIPPE _ Yes of course All in the Family! Family Ties was the one with Michael J. Fox. Ok. Let me just make a quick intro and say a few words about myself. I'm a molecular biologist working in Organ Transplantation. And after majoring in Biology I made a masters and two unfinished PhDs in nanotechnologies and nanosciences. Nevertheless this gave me the opportunity to be part of the development of a couple of biochips nanosensor devices for the detection of viruses. And this scientific and technological background is what gives me the material to address these singularity issues. They fascinate me because of the linkage they make between bio and non-bio, living and non-living stuff.

TOM _ Let me just ask you why didn't you finish those PhDs?

FILIPPE _ Yes, well. Long story short during the time I was making the first PhD I was also working in the Lab in Transplantation, and at the same time I was teaching at the University, so that was all a bit too much. And then I tried it a second time and got almost to the finish line (laughs), but yeah, well, my second chip, well my first one was actually quite successful - it works, it's also a proof of concept. But the second one is actually a lab-on-a-chip kind of device, and due to not having that much time like my colleagues

only focused on their devices and PhDs, so I took about two to three months just in doing, in the fabrication of each device whereas everybody else was able to do this in just about a week or ten days or so. So in the end I gathered too little data and then I thought this is going to take too much time. I'll be ready in ten years or so and by then the chip would already have been surpassed... So

TOM _ Let me point out. I understand that problem very well! I didn't get my disciplinary PhD in Medicine, Philosophy and Geography until I was fifty! (50). And let me just suggest to you that there is a value to a good PhD and you might look towards time to get it done well. Don't close that door for you. I mean you're younger than 50! So you're probably ahead of me. And also the dissertation was on the US engraftment in transplantation system. And the dissertation was published nine months after the defense. Published by a New York publisher. So I'm somewhat familiar with the issues regarding transplantation.

FILIPPE _ Yes. So we should do another interview regarding transplantation!

TOM _ Yes, well let us move on to what you want us to talk about today.

FILIPPE _ Yes sure! Anyway this one's really going to be finished. I have to be done with it I guess by Fall this year.

TOM _ Well the corona virus forcing isolation will help you!

QUESTION I

FILIPPE _ Yeah. Definitely. Ok having that said I think I'd start right into it. To the questions. So regarding Whole Brain Emulation and Mind Uploading. Would you think of another me to be me? Or just a copy of myself?

TOM _ Would a copy of you be another yo?

FILIPPE _ Yes. I mean how would I know I'm me? Or just the copy...

TOM _ Yes that actually comes up in the current Star Trek Picard at the end of the first season. Where Picard is dying and is transferred into this nanobody and given a new life with the same memories but in a different body. Eh... Would they 'clone' me and make another me?! With all my memories and everything I know, that would be me! If they'd

put it in a body of a twenty year old, they would put an old man into a young man's body, and the two personalities would differ.

FILIPE _ Yes but actually the mind upload would only exist in cyberspace, in this case. Yes, imagine WBE will actually be feasible...

TOM _ And in that environment would I think of myself as a material being or would I be just a spectre of a disembodied thought?

FILIPE _ (pauses)

TOM _ If up in cyberspace, the me that is translated, will I think of myself of just *thought* or will I think of myself as having a body in an environment?

FILIPE _ Well if you think that things should go the way that they seem to, there would come the time where WBE would lead consequently to making copies of our minds in a digital form. Of course we would have blueprints of our minds and everything that goes with it, although I'm not discussing at this time if they'd have consciousness and personal identities, but they would be an actual copy of everything that goes on in our brains.

TOM _ Science Fiction has two models here. And they're very very different. One is the matrix model. Where in fact one lives in a cyber-reality without even realising which of you is interacting with the environment and with other people and living experiences and new thoughts and new relations. The other one is an older cyber model in science fiction, in which the brain is uploaded but has no sense of materiality. So there we have the people from an earlier science fiction such of Gibson? In which the mind becomes something that may be able to interact on the screen and maybe able to interact in some cyber way but has no sense of the materiality at all. So there are these two different models each of which result in very different *yous*. One in which you think that you're in fact a whole person and the other one in which you know that you're a cyber construct. In the one where you think you are whole, that is you but a you that is changing rapidly in relation to the environment you're in, rather like a virus that has mutated and becomes something else without really knowing its former form. The other one is basically one where the you becomes a cyber entity without materiality and probably without any way to interact except through the digital to the real world so you would become a ghost of yourself looking for a way to inter-react with the world that you have left. I'm not sure that either of these really... Let me start by saying I don't really believe that these are anything

other than purposely philosophical questions. I don't think we are coming to a singularity, I think that this work is an excuse for not dealing with very pertinent questions that are excellent today. I think that the idea of the singularity is one of those thought experiments which is taking out reality and that neuroethics is basically one that avoids the complexity of today for what it sees in a *futurity* of tomorrow. In some ways I object to the whole line as anything but frivolous where one actually expects this to possibly be a future. On the other hand, given the fact that you're trying to get the damn thesis done I'm willing to play, as long as you understand....

FILIPPE _ Yes but it's ok, we'll get to the singularity shortly.

TOM _ Yes I wanted to put that prejudice upfront!

FILIPPE _ (laughs) Yeah yeah.

TOM _ Go ahead. I think I answered this question. For one it doesn't matter because you think you're living in an alternate world in which you have physicality and in the other one you might converse with your physical self but you know exactly what you have lost. This time the personality will differ markedly from whatever it was at the moment of the upload.

FILIPPE _ Ok. So inside this question I have a sub-question. I think you're going to like this one (laughs). What would represent the biggest or danger in achieving something like this. Making an upload.

TOM _ In either situation the risk would be that the digital version finds itself to the extent that it knows what it has lost. It's one where it sees itself as a *shamira* and will resent it. To the extent that in the cyber world it may become a threat. The other risk....

FILIPPE _ That was actually explored in the black mirror episode.

TOM _ Yes. When I uploaded and found that it was just a cyber person that could not eat chocolate cake and I would sit there and talk to the physical me eating chocolate cake I would turn to me and say "You *schmuck*! You tease me!". The other thing is that there is a long tradition that argues that mortality, that the very nature of our existence has more to it in finite beings. Much of that contributes to the value we place in life and on the immediacy of the day. And I think that I've never myself known whether mortality was one of God's created inventions or simply another example of non-union labour by

an apprentice. But I do think that that is a serious concern. In the nineteenth, eighteenth and seventeenth centuries people were less afraid of being dead than they were of dying. People like Hume (David Hume), knew that they (and the great debate between Hume and Boswell), Boswell believed in the afterlife and with concern for his friend, not his death, but his soul. Hume, the great materialist, had no concern for his soul but also said that dying is a natural and inevitable thing, *Bosy what are you worried about?* So when we begin to posit these things, we posit these as an answer to our fear of death more than our desire to preserve the best of a person. Would have Albert Einstein been *Albi* if he wouldn't be able to play his violin? The violin was in many ways a signature of a kind of thinking that he brought to his work. So I think one of the deep questions in proposing this as a solution to what may not be a problem but a fear, we in fact attack one of the very essence of the things we're used to give life meaning. And as one who has worked in climate ... for many years time again I've seen people come to grips with the process and then in ?belivement? They don't really realise that or take a while to realise that death has been a part of their own life. And in a way that it is fundamental and they have to come and appreciate it. So were these things other than fantasy for people who don't want to deal with the real world issues? I think that's a fundamental question which has to be asked as we propose this as some solution to what is apparently not a particular problem. And frankly, how many of the really frivolous rich do we want to have preserved. The final problem is really who is going to be preserved? It's going to be the Kardashian's of America, that group of intellectually disabled rich people who make money sitting there fighting on screen over things nobody really cares about. I can see all the Kardashian's uploaded into eternit that would be able to interact and basically polluting the cyberworld with nonsense. I think we should have alaw against the upload of all kardashians!

FILIFE _ Or maybe there would have to be some kind of criteria to be eligible to be uploaded... you would have to undergo a quiz or something to be able to be uploaded...

TOM _ As you know from organ transplants that doesn't necessarily work...

FILIFE _ But you know why I'm saying this?, Because the Kardashians also have the right to vote!

TOM _ (laughs) Yes I know that's one of the real stresses of democracy...

FILIFE _ (laughs) Yeah, well, that is democracy, yes!

TOM _ In the eighties, Larry Hagman, who had cirrhosis of the liver...

FILIPPE _ Larry Hagman? The actor of Dallas? “JR”?

TOM _ Yes the actor. Yes he had a liver transplant because of cirrhosis. And if he would not have been Larry Hagman, dubiously he would have gotten that. And he also had some gulp stones, they took the gulp stones for free and he made them into a necklace or something. So he clearly was given the chance of transplant because he was a rich actor (of the b-minus level). So whether or not we can set up the criteria that would be a whole to this question and effective... and again how we can set up those criteria of who will get it will always be questionable. And it's certainly not a question the cyberneticians who are promoting this are looking to with some intelligence. Will we deal with people who are looking for a millennial future, they often don't want to deal with details and haven't really sensed the first two chapters of the New Testament.

FILIPPE _ Ok. So moving on. To the 2nd question.

TOM _ Yes I hope so.

FILIPPE _ (laughs) Yes you won't like this one either...

TOM _ (laughs) I hope you're finding this interesting...

FILIPPE _ Yes of course. Sure.

TOM _ Go ahead.

QUESTION II

FILIPPE _ Regarding Superintelligence (artificial general intelligence undergoing an intelligence explosion of some kind), do you think that when it happens, if it happens, assuming it would happen, and we're still humans without those big enhancements, would we even be aware of it? (clock bangs...) Sorry, it's the old clock...

TOM _ I love clocks. I'm a fan of clocks and time.

FILIPPE _ Yes, me too. More of time than of clocks.

TOM _ Can you see? (shows a clock). An 1865 pocket watch. Go ahead.

FILIPE _ Ok, will we even know it, be aware of it? Can we grasp the moment that AGI actually surpasses us? Or do you think it's never even going to happen?

TOM _ Well what do we mean by intelligence? How do we basically define it? Is it an emotional one? As well as an after .. of intelligence? This was the great problem put forward by the character Data from Star Trek, the next generation. Superior physically and in many ways mentally but learning the emotional before. I would much rather have, eh, some of the people I worked with have an IQ over 140, very very bright people but totally unsuited for almost any place in society, outside of the *Mensa*, where they sit and moan about the fact that people don't appreciate them. Then you have people with Down Syndrome, who have much lower IQ by the standard but often a very high EQ (emotional intelligence). They really get .. about feelings. So, all of this talk about a superintelligence never deals with an abstract of intelligence without an emotional and an historical context would to me be no intelligence at all. It would basically be this vast construct of the cybericians, who might recognize that there is no AI. There is intelligence, however it comes in terms of an ability to read and collect and understand and interpret data in the light of the theory and the history. Machines don't learn. Machines are trained through very precise algorithms to very very specific tasks.

FILIPE _ What about Machine deep learning?

TOM _ Well it's not deep learning! It's deep training. One man teaches a very powerful computer to play chess logging into a history of several hundred years of games and rules. And from that the program can extrapolate based on all the movements in time. Is that intelligent? Well, it's clever. But will it allow that deep brilliancy? Maybe not. Outside the fact that it's imitating the brilliancies of the past. Can it then go off and take that and be strategically useful and look to the health resources during a pandemic? No that's not its job. It is so narrow and so specific in its programming while it's interesting for chess and in the chess world, it is too narrow to be thought of as an intelligence. Deep learning is really deep training in a very specific area. And there's a vast difference we have to recognise even when we think abstractly about this future nonsense. I suspect we will end our race well before even the possibility of something like this coming forward. Then we'll have to leave it to the cockroach evolution will take over in 50 million years...

FILIPE _ (laughs) Or more. Ok, but even if you don't acknowledge (I would use this term - acknowledge) the possibility of Artificial Intelligence, actually existing and progressing

in time, think of it as a thought experiment. If another intelligence, beyond ourselves, our own human intelligence, would actually surpass us, and just a few seconds afterwards it would be so advanced we wouldn't get it! We would have lost that track...

TOM _ I don't think that's true. I mean first of all i'm not denying the possibility so much as denying the definition. What is the definition of intelligence? How do we take, if we think of intelligence as an abstraction, with emotion in the history of physicality. The problem of this literature is that it doesn't address the whole nature of what they mean by intelligence as an embodied concept. I think it is begged to ask what would happen if another race appeared which is so far beyond us in technology. I would point out that we would understand to a certain level. One does not have to be a music expert...

FILIFE _ Yes just let me, exactly in music you know the examples of AI nowadays, and we're still speaking at this moment of a narrow AI which is quite different from the so called General AI, that is a kind of human level artificial intelligence. So we now only know the narrow one, and this may be conditioning our thinking and reasoning. So anyway, you already have nowadays examples such as narrow AI able to complete the tenth symphony of Beethoven, for instance...

TOM _ No. To provide a possible ending of Beethoven's tenth symphony.

FILIFE _ Yes but you were talking about emotions right? There was an experiment, they're making so many experiments in this area now, but, did you read *Home Deus* by Yuval Harari, by the way?

TOM _ No.

FILIFE _ Ok, butso I read a few interesting examples in his book. And one of them was exactly regarding music. There was a composer in the US, who was also challenging AI so to speak, and said "No c'mon you can't make music such as Bach did, you have to have emotions, you have to feel whatever... So he then got into this challenge with a friend of his. And the challenge was like "Ok, I'm going to make via AI composition. A music. And then I'm going to show it to you and your students, and then we're going to play out your music and another one by Bach. I think it was Bach, I'm not sure. So they made this event this day full of people, a big audience in college and with 'ordinary' people too so to speak, and the result of it was that the majority of the people thought that the composition of the AI was actually Bach! And that the music played and

composed by the teacher of music science was actually the one made by AI! So the majority of people got this feeling from that music and hey look at that, it's actually made by an AI! And we're still talking about narrow AI!

TOM _ And here exactly is where it is not intelligent. Yes you can take Bach's form. (phone rings)...

FILIPPE _ Someone's calling you.

TOM _ Yes, I'll let it go. If it's important I'll pick it up when we're done! The genius of Bach was in his creation of a music which had not occurred before he was advanced of a whole series of musicians. To a point where Mozart's genius wasn't simply an improvisation but in his ability to integrate what has been something, and in timing which had never occurred before. In the late nineteen and beginning of early the twentieth centuries colleges were beginning to listen to african music. They really couldn't understand most of it. Even the time or some of the notes. Between F and Fsharp there is something that the african singer can hear that we cannot. Aretha Franklin can sing it most of us can't even hear it. Matt Schulman writes about this in the history of early Jazz and I think that Charles Isac was one of the few early people to understand the very complex time that was there. Is that basically an advanced intelligence or a different intelligence? It's certainly basically a different musicality. So when you have a computer which looks at a lot of Bach. And Bach wrote over 300 pieces. He wrote a lot of stuff of forms. Some of them were genius.

FILIPPE _ Most of them, for sure...

TOM _ But they call it in the english the "?!". *Ich steh mit einem Fuß im Grab*. I stand with one foot in the grave. A small little piece of my larger one. Still it means so simple and evocative. One can copy it. One can improvise on it. One can change the signature. But it remains one of those unique pieces in the history of all music, which is universally accessible. What you call Artificial Intelligence I would call basically a narrowly trained machine can replicate the basics of the form can give us something very likable and advanced in music but will it say how can in fact take this to my male voices singing how can I change this. In London I was in Saint Patrick's church. And there was an argument in the new composition based on the Gregorian form. I thought it kind of uninteresting until halfway through when I began to hear what sounded to me like diminished and augmented chords all the time. This was astonishing. So afterwards as everybody was

leaving they said why did we change in the middle of this to what sounded like an augmented and diminished chorus. He said the organ is really bad prepared and two keys are out of place in the middle of the play. So what he had to do was to change the structure of what he had written into a way where he could get around this failure. I don't think any of these programs are of the nature which would understand music so technically I know none of them is recreating a bass sonata? But in fact we are taking the music and moving it forward in the way that we find some of the existing in the twentieth century something that deals with those forms. So I don't think this is intelligence at all, I think they're basically just better at copying.

FILIPPE _ Well I was giving you this example but I could give you lots of ...

TOM _ An excellent example for the problems with the language and the problems with assumptions that are undergoing on this on what is intelligence and what these technologies can and cannot do and will we in the future.

FILIPPE _ A couple of years ago, two or three, there was a contest of sci-fi short stories with I don't know four or five thousand people or so. And the story that won was actually written by an AI and it was ironically about an AI that felt trapped in itself.

TOM _ That implies a level of consciousness that we cannot at the moment expect...

FILIPPE _ Yes, and remember we're still talking of narrow AI. I agree with you, even with deep learning machines it still is essentially programmed. But I have to pose, even as a scientist if you will, I have to pose the possibility that there are hypothesis that maybe in time and progress as it goes, exponentially as it is going, there could be a time where it basically could or would surpass us. We have to imagine this possibility and the real question here is if it were to happen, you could also pose this question as another kind of thought experiment and say that it is a biological extraterrestrial kind of intelligence, right? So we humans would at that time be probably unaware of it. I'm going to make a comparison with the Fermi paradox. It's not the fermi paradox? The one where you won't find extraterrestrial intelligence, even though the probability might be very large, but simply by the fact that it would have to be a kind of intelligence exactly at our level. Because if it's just a little bit, one second in time of evolution of its chemistry, its physics, its biology would make all the difference. So it's like going to Mars to watch bacteria but ok we won't communicate with them and they won't even know that we exist and we

won't bother in contacting them. So this is the paradox, and we won't find intelligence, unless it is exactly at our level, but for that they would have to be us!

TOM _ No I don't think that's true. There are all sorts of ways to see that that's not true.

FILIPPE _ Well, that's the paradox.

TOM _ In music people didn't understand what was going on in the african musical world, the early musicologists. But they knew that there was something there. They could perceive it but didn't have the tools to be able to understand it and describe it. And right up to Paul Simon and his work with the late black Mombassa, the rhythmic .. of the african music was something which almost nobody else understood. Bacteria and viruses I would describe as mindlessly intelligent. Certainly they know about us, they interact with us in a way which is profound for their evolution. I did a whole lecture in one point: the politics of the virus and why the viruses will win. Because they're socialists and we are not. Viruses share "Hey look at my protein! You'd like some? Sure! Try some of mine! Oh the red one is cool!". We on the other hand react not at all until they become a threat and then study the genotype and figure out a way to find a vaccine, make some money, submit it to trial and then we're at the great great grandson of the original virus. So yes we could find ...

FILIPPE _ But you think they do this consciously?

TOM _ No, I said they are mindlessly intelligent. Would we have visitors from outer space, say Artur Santuris. A race in a spaceship, clearly technologically advanced. Then the great question would be what advances did they make besides technology. And would say we would be able to see it. Would we accept it? Look, the whole point of the far right movement in most of our countries is that there are people that will not accept the lessons of world war II and the globalization we have today and think we should go back to some kind of nineteenth century armed individualism. So do they perceive the great technologies of humanity? No. Would we perceive it if our face would be rubbed into it? Sure let that big spaceship come. Will it come from computers that say that we'll be the ubermensch? Nein! And that won't happen as I said a type of disembodied type of intelligence in thinking in terms of a world which is basically rooted in its physicality and its mortal history.

FILIPPE _ Ok, Let's move on..

TOM _ Yes I will have to see at some point what this call was about...

FILIPPE _ Oh, ok you can if you want...

TOM _ No let's finish this first.

QUESTION III

FILIPPE _ Ok. Next question has to do with transhumanism which I gather you dislike a lot. (laughs). So imagine in a few years from now, because of human enhancements generally speaking, it doesn't have to be mind uploading or whatever, I mean I could also say I'm enhanced at this moment simply by wearing contact lenses right? So in a few years from now, there will be because of technological progress there will for sure be available a bunch of mindblowing enhancements in our bodies in our minds whichever. So this possibly will split society in terms of who gets access to them and who gives access to them. I mean is it the government? Is it google? Is it Trump? Who is going to give access to specific enhancements and the question actually is, a few years from now politics could be divided between bioconservatives against transhumanists.

TOM _ Did you see my paper on transhumanists and those 76 trombones?

FILIPPE _ No. Not that one, no.

TOM _ I will have to send it to you.

FILIPPE _ I read "Enhancing who? Enhancing what? Bioethics and Transhumanism"

TOM _ I recently had a paper published I think .. called Transhumanism and those 76 trombones. ..was the music man and the false promises. I will not say you're enhanced. Certainly you are beginning to go bald...

FILIPPE _ (laughs) Not yet!

TOM _ You could have hair plugged in. I have two prosthetic hips.

FILIPPE _ You're enhanced in a way...

TOM _ No. There's a difference basically between prosthetics and enhancements. The very interesting work which has been done on creating artificial limbs which can answer

neurologically to the brain's movements. The ability to actually begin to create a link to people who were locked in, or say in advanced ALS to be able to direct thoughts returns them into a physicality into the world in a prosthetic fashion. If you have a mechanical arm where you can actually move your fingers although you cannot really play Bach. That is not an enhancement. It's a prosthetic. Your glasses, my glasses, contact lenses, a hearing aid, is not an enhancement although it may improve one's quality of life.

FILIPPE _ You know that spanish guy (catalan-swede) that has an antenna attached to his head, he basically sees more colours than we do. Would you think that's not being enhanced?

TOM _ Well I think that's more prosthetic. My friend...

FILIPPE _ But he does see more colours than we do. We see more colours than Lions, they only see sepia. So we are kind of enhanced in relation to that...

TOM _ There are also those people with different neurological conditions. Remember that from what we know from linguistics and is that we know what we can express. The ?? have like (15) fifteen words to distinct types of rain. There are in other languages words for different snows we don't have words for we don't perceive those differences. The old .. chinese physicians and tibetan physicians hear twelve (12) beats to the pulse whereas we normally hear at the most four (4). So some of this is what we have learned to name and perceive rather than necessarily with the physiology. So for a lot of these things it is about not an enhancement but a prosthetic. And an ability to name them and to integrate them in what we can express and utilize. So again I find it different in more prosaic interpretation here than you do.

FILIPPE _ Yes but that's linguistics.

TOM _ But it's linguistics that allows us to express.

FILIPPE _ Yes, I understand you, but in political terms it may really have importance here, and you know this better than I do, but nowadays in therapeutics I swallow pills, I take stuff for migraines etc... So the thin line between therapeutics and what I call enhancements and you call prosthetics is that it is a thin line, and the basics is that you have so many and they all cost money of course and you don't have a health care system in the US...

TOM _ Remember I'm in Canada...

FILIPE _ Sorry! Of course, oh I'm so sorry. No, you have one of the best healthcare systems in the world (laughs). So true. Baad example. But you know what I meant with this. Some people will have access to specific things, I don't want to delve too much into linguistics... You do use a smartphone right? And I would think that there still are people in the world without smartphones. In ten years from now some of us probably can engrave those smartphones inside our body hence I'm enhanced! You could call it a prosthetics but ok, I do have an advantage, an adaptive advantage because I'm online all the time with google, and google is inside my head and the guy from Burkina Faso doesn't have a smartphone inside his head. So what I mean with this all is that there will be sometime where the world could be facing a kind of dystopia where people get *enhanced*, in italics, I'm just going to say *enhanced* in italics, and then there are those that don't get enhanced... So this is going to divide politically...

TOM _ Well I don't call them smartphones. I call them basically digital cell phones. They're not smartphones.

FILIPE _ laughs. Yeah.

TOM _ They're not smart. We are. And I would point out that one of the interesting things in epidemiology is that, and I was one of the first to write about this, is that the ability to use them with global positioning in epidemiology in outbreaks far in some parts of the world. You don't have to look to futurism. Will the person in Burkina Faso be able to get one of the new prosthetic limbs that some of the Americans are getting? No. Will the DA provide every soldier that has been injured with multiple amputees. No. What we already have is a dystopian divide in every level. In our cities in our countries and globally between the rich and the poor that the advantage is available to each. I think that's impertinent by trying to place us in the future in which we will debate this takes us away from the fact that we are in that present in basic care and basic education. It is inconceivable that we can, that is you and I and in the richer countries, access webinars and resources which people with greater need in Burkina Faso or the democratic republic of Congo might dearly need but might not be able to access. When I was young I worked in clinics in Guatemala and in Ecuador. At best we could do nineteenth century medicine. But that was a grand advance. Of what was then available. So I see no need to put this in the future that I don't believe in when I see this happening in the present. And I see

putting in the future is a way to not deal with the inequalities which are rampant here and rampant in the introduction of new life preserving and improving technologies for all.

FILIPPE _ Ok. I got your message.

TOM _ My friend Fabrice ?? who is an american bioethicist trained in philosophy.

FILIPPE _ What's the name again? I didn't catch the name.

TOM _ Jartrand? Fabrice Jartrand? He's swiss. I'll send you his email. He teaches in the US. Fabrice Jar...

FILIPPE _ Ok... Well you know a bit of portuguese you know spanish and german and your french isn't that... And you're canadian living in Toronto! (laughs)

TOM _ I used to make translations in english - spanish -japanese and chinese.

FILIPPE _ Japanese and Chinese... Ok

TOM _ Ja, und Ich bin ein Koch

FILIPPE _ Yeah yeah

TOM _ Deutsch! Et je suis canadien et parlait française du Québec!

FILIPPE _ Du québec! Hahahahaha. Ok

TOM _ (laughs) Fabrice is making a big name of himself in neuroethics. And I said to him why do you deal with this stuff when there's so much else?" It's a way of not dealing with these medical issues. Fabrice's answer to me was that people are writing about this, about these technologies that are coming, we need people to discuss these eventualities. My answer to him is I'm all for philosophers doing inconsequential things. But I think that in terms of ethics we need to recognise that these are in many ways inconsequential in the immediacy of the day and that the problems that you posit to the future are concrete realities of the present when we understand the limits, when we define intelligence strictly and look at these problems as one of the inequalities of services of today with the enhancements which aren't that advancing. We made extraordinary advancements for instance in understanding attention deficit in dyslexia. In the way that dyslexia can be trained. I remember in the early nineties students that had terrible troubles who once they had a computer could in fact work in a way they had not been able before. Was that

artificial? Was that intelligence? No. That was basically a work around for a type of perceptive condition. Is it a disorder? There are certain things in many of these neurologies which are wonderful explorations. I have no spatial pattern. Kind of funny for a map guy. If I walk into a building from a different entrance I am lost. I have to always enter through the same entrance to move around. I quite literally lost one night. A spatial inability. A person with autism has a spatial programming ability. Many of them I can dimly understand but I cannot achieve or basically hardly perceive. So in the diversity of the human condition, there are those with differences which may in fact be attributes we should recognize them and I say there is where we look for basically intelligence we seek in society for the people we need to help direct us to the future not to the computer deep training which has many benefits but is not to me intelligent.

FILIPPE _ Ok, Tom.

TOM _ So I've ruined your whole god damn project and now you're going to study neuroscience and become a neuro?? I'll send you an email with my paper on transhumanism and I'll also send you the email address from Fabrice who is deep into this stuff with some skepticism but in a level of acceptance I don't really buy. But he'll maybe be a useful person for you to talk to. And I hope you found this interesting.

FILIPPE _ Na Klar! Sehr interessant.

TOM _ Merci beaucoup.

FILIPPE _ It was different. An alternative vision, but I value it as much as all the others as I told you before in our email exchange it's important also to me to exactly make something out of this that is kind of neutral, if you will. This is an analysis, and an analysis has to be critical too, otherwise it wouldn't be an analysis...

TOM _ I think it is important in this, and I congratulate you on it. To have not only those super salesmen of the future. But also the critique of them. I think I stand as one of those who presents his critique.. I've done it in genetics. And I also have done it with the transhumanists. And with advances we have to understand that thought experiments are of great value if they tell us about the present and not to ignore the present for the future. Which will probably never be. Or will be in a way that we will not conceive. I would like to see those people from artur sancturis come down and talk to the rednecks in oklahoma!

FILIPPE _ (laughs) Yeah, the rednecks in Oklahoma...

TOM _ There was in science fiction I forgot where it was which basically said you people can do whatever you want to but we're going to stop wars. And you can't hurt animals. And they immediately put an end to all hunting. No big game hunting. Subsistence was ok but no hunting. The innuick would go and do what they've always done but the rest of it no, and they stopped the bull fight so that anybody in the audience could feel the pain of the bull. I like that. My kind of intelligence.

FILIPPE _ Sure!

TOM _ I hope this was helpful.

FILIPPE _ Yes of course.

TOM _ I hope your chips make a lot of money for you and your PhD that you didn't have time in the past. Simply because a good dissertation teaches things that nothing else does. And the title does give you a certain standing. Let me just finish that I have a multiple interdisciplinary PhD of the University of British Columbia in Medicine Ethics, Philosophy and Geography. And I work in four fields of these areas. Let me know how it goes. I'll send you that stuff in the next day or two and let me know if you'll need me again.

FILIPPE _ Ok! Mache ich!

TOM _ So you're in Portugal now?

FILIPPE _ No. Ich bin in der nahe von Stuttgart.

TOM _ Ahh Stuttgart! I had an old girlfriend there. But I'm sure she's a grandmother now.

FILIPPE _ (laughs) Ja meine mutter ist von hier aus Stuttgart.

TOM _ But your university is in Portugal. Porto or Lisbon?

FILIPPE _ In Portugal yes. Porto. The University is in Porto yes. Although I live in Lisbon

TOM _ I'd love to visit it someday. I am told it's a very civilized city.

FILIPPE _ Which one? I mean both are! (laughs). Lisbon is a beautiful city.

TOM _ I'm coming to watch your defense!

FILIPPE _ Oh I have to be careful with that (laughs). But yes, Lisbon is very nice and everyone should go there once in their lifetime to visit.

TOM _ That's what I've heard yes. Ok Bonne Chance

FILIPPE _ Ok thanks. Tschuss!

TOM _ Auf Wiedersehen!

THE END

APPENDIX J INTERVIEW WITH JAMES BARRAT

(Filmmaker / Author, USA) [skype-60min @ 30MAR2020](#)



FILIPPE _ Hi. I think now you can see me right? And recording...

JAMES _ Ok you're not planning to broadcast, right?

FILIPPE _ No no. It's only for academic purposes. And I have to have the digital backup...

JAMES _ Sure! Ok.

FILIPPE _ Otherwise I could always say that you said this or that... yeah but show me the file, you know?

JAMES - Yes. Sure.

FILIPPE _ Maybe the Jury will watch it... But that's it.

JAMES _ (laughs) Yes that's fine!

FILIPPE _ Ok. So a quick intro about myself, and as I told you on our email exchange I'm a biologist and have been working in organ transplantation for almost twenty years now.

JAMES _ Fascinating.

FILIPPE _ So after majoring in biology i made a masters in biomedical sciences and then went on to PhD on nanotechnologies and nanosciences, and so this gives me the scientific and technological background to address these issues, these singularity issues, hence this new PhD, well it's almost ending not that new anymore and these interviews that I already made to some people like yourself, and basically that's that

JAMES _ So this is your second PhD?

FILIPPE _ Yeah...So having that said and because I don't want to take too much of your time

JAMES _ Go right in!

FILIPPE _ We can start by... I read your book, by the way. I like it a lot, of course otherwise I wouldn't have contacted you. And as I told you I'm not in favour or against it i just think that there will be some feasible technologies that will actually change our human condition. Everything! They will change everything. So I got 4-5 questions and so:

QUESTION I

FILIPE _ Regarding WBE and MU, would you think another me to be me? Or just a copy of myself?

JAMES _ That's a very good question. I'm not a scientist. I think a lot about that. And it's a very good question, I think that you would be just a copy of yourself. I don't know how you would transfer, you know, the essence of someone to another me, you know. Ray Kurzweil proposed that we'd do some kind of nanotechnology approach where we would take the brain cell by cell but it would destroy the original brain in the process. That copying of the brain cell by cell would somehow transfer our essence or soul or whatever you want to call it. I'm very skeptical that I think at best it would just be a copy. Just because we don't even know what we're talking about when we say things like soul or essence or our consciousness. We don't know what it is. I once thought about writing a book about consciousness. But then, you know, consciousness is like a puzzle. It's so complex and there are so many ways to look it at. The idea to transfer that essentially to a machine or some other substrate I just think is preposterous at this time. I think we could make a copy and then to me the whole, maybe we'll get into this in other questions, but the whole mind uploading and, you know, what it's really about is about living forever. And in my opinion, transhumanists, like everybody else, have a deep fear of dying. And this is in my mind, like one of those created religions, with religious promises that will never die. And naturally as biological organisms we're afraid of dying. It's our prime directive to stay alive, and procreate, is our second directive. So we're desperately afraid of death, we want to live forever, and so we're creating this mind uploading idea. When I read Ray Kurzweil's *The Age of the Spiritual Machines*, I was very much in favour of it. I thought this is mindblowing...

FILIPE _ Mindblowing, yeah...

JAMES _ I remember I was euphoric about the potential then. But then I became sceptical, and not just of the ability to do it but the motives behind the people who want to do it.

FILIPE _ Yes, well those aren't the only motives behind. There might be some other advantages. But you said, you were talking about consciousness, and without knowing what it is, nobody does, nevertheless everybody thinks they have one, right?

JAMES _ (laughs) But they could all be faking it except me!

FILIPPE _ Exactly, yes. But you know we could emulate our brain capacities even without fully understanding them. Technically it would be feasible. So the fact that we don't know what consciousness is, doesn't mean we wouldn't be able to transfer it.

JAMES _ Right. We could, but we could end up making just empty vessels, you know? Empty emulations.

FILIPPE _ Of course yes.

JAMES _ We could make avatars.

FILIPPE _ Yes. Exactly. So next question, actually a sub-question. What would represent the biggest risk in achieving something like this?

JAMES _ Well, here's two risks in WBE. One, emulation can lead to augmentation, which can amplify the ability to whoever that gets augmented. And as I said in my book "Our final invention", the people who will get augmented or who will get the brain enhancement won't necessarily be the most ethical. They'll be the most rich ones and as a study I quote in my book says that richer people aren't actually nicer than normally average people. (laughs) In fact they are meaner, more selfish, they take more and a certain percentage, and here's the real danger, a certain percentage of our population, including a lot of politicians and CEOs are psychopaths. Our president of the US is an example of that. He's a narcissistic psychopath and we only found this out by looking him under a microscope, which we're doing now and every day and it's making us all ill. But we see that upclose, that narcissistic psychopathic politician and businessman. This is the kind of guy most likely to have his intelligence augmented. So whatever bad things he's doing will be radically magnified. He could do many more bad things at a faster rate and outsmart us all with that super brain, and he'll live forever! So to paraphrase Elon Musk we could be spending our lives governed by an undying narcissistic psychopath. And that's about as dangerous as it gets. The other big danger, besides letting psychopaths take over, is if everybody lives forever what about the new people being born. We already have, I think by 2040 we're going to have about 10 billion people on Earth. We have no plan on how to feed them, and yet the population continues to grow. If you have a certain segment of the population that's living forever, using resources, and this is a side-effect another implication of brain augmentation and brain uploading, if you have a bunch of

people living forever you're by necessity excluding a whole other bunch of people from living meaningful lives that are running out of resources. Again, this will be the rich people using a lot of resources and again leaving less to the poor people who already have a vast economic gap between the richest and the poorest on our planet. What this does is, it invites it to get worse, faster! So I don't think we should solve the problems of immortality but we can solve the problems of mere mortality so that fewer people die every year. Take the city of Chicago and it has something like 8 million people. Imagine ten times the amount of that in people, that's the number of children who die of starvation each year. So in ten cities like Chicago how many die of starvation each year? We have to take care of them before we start making a class of immortals.

FILIPPE _ Yes ok, but the uploads could live literally in cyberspace.

JAMES _ Look at the carbon footprint of our artificial network. It's astronomical. Cyberspace does not take zero resources. And these people would be consuming resources and demanding services and doing other kinds of things rich people do. That wouldn't be comforting. Cyberspace is not a zero resource allocated.

FILIPPE _ Ok. You were talking about politicians and psychopaths, CEOs, psychopaths, but you know that's a bit of the history of the World, right? I mean basically we're ruled by idiocracies.

JAMES _ laughs

FILIPPE _ Yes, it's true and I think the democracies tend a little bit to this path, you know towards this idiocracy.

JAMES _ I think it goes in waves. We got some really good people and some really bad people. I think the pendulum swings both ways. You someone like Barack Obama and then idiots are outraged by that and the pendulum swings back to someone who's a nationalist like Trump. And that happens in every country. You got Boris, what's his name in the UK?

FILIPPE _ Johnson.

JAMES _ Yes. And then the pendulum will swing back and they'll have someone good again.

FILIPPE _ You have bolsonaro in Brazil too and... He's actually a copy of Trump in many ways...

JAMES _ Yes. It's incredible how these people are the ones who rise to power.

FILIPPE _ Sure. But then again on the other side you have...

JAMES _ You have Merkel, which is great!

FILIPPE _ on the other extreme obsolete governments like China...

JAMES _ Oh China is, well it doesn't even pretend to be anything else.

FILIPPE _ (laughs) Ok. Moving on to the second question.

QUESTION II

FILIPPE _ So regarding Superintelligence (artificial general intelligence undergoing an intelligence explosion of some kind), do you think that when it happens, if it happens, assuming it would happen, and we're still humans without those big enhancements, would we even be aware of it? Imagine AI would in fact undergo an intelligence explosion of some kind, and progress in such a way that in a few seconds it would surpass us and we wouldn't even be able to understand it, to be aware of it?

JAMES _ You know that's a hopeful and very good take on intelligence explosion. How would we, as Stephen Hawking said, superintelligence would be able to create weapons we wouldn't even understand. In fact we wouldn't even understand anything it did. If somebody or something is like a thousand or a million times more intelligent than we are, in a mathematical rational sense, not in a musical sense, dance sense or kind of aesthetic sense. But in a mathematical rationalist sense we really wouldn't understand anything about it. It would be like suddenly sharing the planet with the aliens. But then the question becomes: Will they be nice to us? How will they treat us? Will they disregard us or will they kill us with neglect? Will they use our resources, you know, we have no idea and no experience of living with something a thousand or a million times more intelligent. And that also amplifies the danger of sharing our planet with superintelligences that are in human minds. Augmented people with superintelligences. We used to think, back when

people started thinking about existential risk and AI in the year 2000, and I'm talking about Eliezer Yudkowsky and...

FILIPPE _ Yes I was just thinking of him because of that quote he has, you know "the AI doesn't love you nor does it hate you, but you are made of atoms it can use for something else!"

JAMES _ Right! So, well we're sharing our planet with something that's... We used to hope, I think Ben Goertzel said "Let's get to an advanced AI quickly so that the first AI can be human augmented, but again we have no guarantee that will be nice either. In fact there is a certain probability that it wouldn't be nice. If it came with our mammalian heritage for staying alive, with revenge, if it came with all that then superintelligence would be very dangerous. Yeah, it could happen and we wouldn't know it. I used to think about when I saw some people writing something about sitting around a Geiger counter if superintelligence would emerge from some complex system. How would you know that suddenly, well they'd probably use a lot of energy so suddenly you'd have a giant energy drain or it would disrupt some critical systems. So how would you really know? And of course we're referring to the hard take-off from intelligence explosion, it doesn't happen gradually it happens very suddenly. I agree with that. I don't see why it happens gradually once we get to AGI, once we get to human level intelligence, then beyond that...

FILIPPE _ It just happens...

JAMES _ Yes, it would take a few seconds. So yeah we wouldn't probably know it. But I think we would know it pretty soon after that...

FILIPPE _ In this kind of progress a few minutes could mean millions of years in natural evolution so to speak.

JAMES _ Yes and that could mean instant annihilation for us. It's hard to imagine something that is astronomically strong and powerful not being disruptive to our climate or planet or resources, the things that keep us alive. Just by neglect, not even by avarice or wanting to kill us for some reason or something just by not paying attention that there are other living things that are biological living creatures. We used to think, wouldn't they be grateful for having been created by us, but then you have to remember that gratitude is a human emotion built up, again, our mammalian heritage and it's not something a machine is going to have unless it's very exclusively ??.. So that's why people aren't

working on that problem now at the Machine & Intelligence Research Institute and The Future of Life Institute, and all the other Institutes that are .. for so many years.

FILIPE _ Yes, I guess the next sub-question we already covered this, how possible would it be for a singleton of some kind to rule the World?

JAMES _ It's all in gamble. I think that Steve Omohundro (computer scientist*), and others (I wrote two chapters about this in my book), he writes about an ecology of AIs that are safer. So you have, instead of a singleton which could be very demonic, and again to paraphrase Elon Musk I think, we would be ruled by an immortal time rate. And we ourselves could be immortal, so it would be like Hell. Steve Omohundro talks about an ecology of AIs that grow roughly in steps together. There's two things. One is a scaffold of AI approach where you make an AI of limited intelligence to help you get to the next level of AI, and then you develop a better one where you create a scaffold to get to the next level, but never take away the scaffold and you always limit its power. That's one way to approach this. The other way, the ecology that a lot of AIs are maturing at the same time, and the theory is that they'll keep each other in cheque. None would grow so quickly or dangerously that it couldn't be stopped or slowed down by the others. But here we are talking about an environment of AIs and it's just, again, a science of wishful thinking.

FILIPE _ Yes I would apply some universal darwinism in that thought because probably those AIs would compete with each other.

JAMES _ Well, they would! If you read my book then you read about Omohundro's basic drives. The idea that AI will have basic drives of self protection, it won't want other possible hostile AIs to thrive in its environment. I like that theory but he's trying to impose some order on this unknowable thing. You know, how they will behave and he says they will behave with these basic drives, because rational agents and economic period dictates this. And he provides proof to show that rational agents will behave this way. However, I still think when it comes to super intelligences all bets are off. We won't know fundamentally how it will behave.

FILIPE _ Yes I could envisage the possibility of something like google turning into some kind of singleton...

JAMES _ Yes!

FILIPPE _ which then could mean a lot of things, one of them being that all of us together in the cloud, would be this singleton, which also would be very strange...

JAMES _ Very strange, yes. I think, if you read *The Age of Spiritual Machines* it's the...

FILIPPE _ Yes I did. But so many years ago... I have more present to me *The Singularity is Near*, which is more recent...

JAMES _ Yes which I didn't think as good but full of data. But the *Age of Spiritual Machines* talks a lot about living in this virtual world with all those brains working and inner capacities, which is a wonderful idea, but again it's an idea that is seeking for a religious paradise. I'm not sure but again I think singularitarians is really just a kind of religious movement.

FILIPPE _ Yeah Ok this leads us perfectly into the next question:

QUESTION III

Regarding Transhumanism and Posthumanism, will being human be different in a Post-human Era than it is today? Of course if it will... and if so, then would this mean we would definitely be more bit than bio? So what I mean by this is we'd be more immersed, more mixed, a kind of cyber sapiens, so not that biological anymore...?

JAMES _ Cyber Sapiens? That's good. Didn't hear that one before. So I'm trying to imagine a transhuman, a post human World...

FILIPPE _ Ok, because this is also leading to another question which is the dystopia, that I can imagine in about 20 years or 30, a society generally speaking with basically two kinds of people, divided between those that get enhanced and get upgrades and those that don't and again, like you said (and me too), eventually this would be the rich people... So the rest of the society wouldn't get those enhancements and you could end up living in 20-30 ys, or sooner, or better sooner than later we could end up in a split society which then leads to the other question of the possibility of politics being divided between bioconservatives and transhumanists...

JAMES _ Wow. Well, back to the first question.

FILIPPE _ Yes! Sorry!

JAMES _ Again, we can't escape the fact that humans have been around for 200 Thousand years, as homo sapiens, and before that we were, well millions of years as primitive humans. We can't escape that, and that baggage is good and bad, it's why we steer the future of our planet and we're steering it into the ground! But it's also the element, well our intelligence has enabled us to survive. I think we can't escape our biological roots, if we have an augmented brain, it's going to be an augmented mammal brain, and the core of that is a two million brain. And we can't, we cannot transcend that! Even when all our needs are answered. We still won't be satisfied. We always quest, we always seek, we're always greedy. We can't stop eating potato chips because we have a genetically hardwired fat seeking gene. We can't stop drinking sugary things we won't even stop stealing things. Part of our society will be brutal and ugly that part will go over to the cyber world just like the good parts. Again, the rich will be the first people. So I think we'll be mammalian, fundamentally, we'll just be enhanced mammalians, I don't think we'll bring out our spiritual nature, I'm not sure why ... of nature. So I disagree with dystopia, you know, why dystopia? Let's look around ourselves and we look at the corporations that are in charge of this intelligence explosion. Facebook is just a huge manipulator of privacy and stealer of data, dishonesty to the core. Google, totally corrupted by its vast bottom as well. Apple uses slave labour in China, among other things. Corporations are almost by definition corrupted at core because they ...

Clock Wall bangs

JAMES _ (laughs) So take these corporations, take their ethics into the future. The future is governed by the cyberworld, and that's a realistic argument extrapolation of what the future looks like. Regarding politics, I think by the time of transhumanism there'll be no politics anymore. I think any notion and representation will be gone. And the monolithic organization will be battling for the control of our pennies of those of us who are left. I mean look at China, look at our country how easy it is to buy politicians. We thought we had a democracy. What we really have is an autocracy, where people's votes and people's representation is just being purchased. Google employs four hundred (400) lawyers! 200 of them are probably on contracts and the other 200 are probably because they get sewed all the time. They are sewed by fourteen (14) other nations for privacy violations and out of theft. They don't pay taxes in the US, they pay taxes in some places like Ireland because...

FILIPPE _ Luxembourg.

JAMES _ Yeah, one of those places. The Bahamas. So I have a dystopian view, but I think it's based on...

FILIPPE _ What we have today.

JAMES _ Yes, what we have today. And my challenge to transhumanists is what are you basing this utopian vision on? What period of Man's history has approached anything like this? Or sustained itself? There have been some, a couple of communes and short.. Of but hasn't sustained culture or...

FILIPPE _ Well, you know transhumanism has actually a historical, biblical background. It's not that it was invented now or a few years ago, just that it's getting enhanced now due to all these technologies and so they support their cause, or religion if you will, with all these technologies now available but the idea is quite biblical. And has political roots too!

JAMES _ Oh yes of course.

FILIPPE _ They're mixing this idea because like I was saying earlier there are going to be, well there already are today, we don't have to step forward ten or twenty years people that are more enhanced than others. I'm more enhanced than the guy living in Ethiopia, for sake of the argument, just because I wear contact lenses and I have a smartphone which makes me more intelligent than the guy from Ethiopia... So, but in a future with this progress and also with the difference between therapeutics and enhancements not being that big, it's a thin line, so there will be a time you will upgrade not for medical reasons but because you need to be adapted to survival, and survival has become something different. So you will definitely have people enhancing themselves and so you have to have in political terms, all those supporters, you know instead of having the classical left wing today I would call it the bioconservatives of tomorrow and on the other hand the right wing the transhumanists. But in dividing this, actually there already is a political transhumanist party in the US, I've seen this on a documentary two or three years ago...

JAMES _ Yeah have you heard about the president of Canada? There's a transhumanist president in Canada, he's just silly. It's just silly. You raised a very interesting idea I thought about as you were talking about it. Genetic engineering right now... I make documentary films. And I made a film, a short film, about a kind of childhood blindness

called liver amerosis? That has been, not perfectly cured, but mostly cured by genetic engineering. They take this virus, take out its DNA, put it in the genetic code for the photoreceptor that is bad in some children. So some children have a bad photoreceptor. They take a healthy version of that receptor, put it in the virus, inject it in their eyes and the virus does what it does best. It invades the cells and changes them, and these kids are getting their sight back. So they're enhanced. Now they're seeing twice as much! Now they are in some sense transhumanists, and that is to me great news for this technology...

FILIPE _ Sure but maybe we won't be able to divide that line in the coming years...

JAMES _ Well what I like about that is that there is a money gap. Right now the only people that are getting this technology, except the people they used to experiment it in China, but the people who benefit from this technology have suffered a disease. And it's being corrected. And I think that's a great application of nanotechnology and transhumanism technologies... Let's hope the people who have it, that they're the disabled. Let's cure some of these, Let's cure all these gene based mistakes. Some of them are being corrected through the neural?, not all of them can, because their bodies are affected by muscle dystrophy. But muscle dystrophy is a terminal disease where you die and you die young. It could be arrested. So that would be, for me, positive news from transhumanist technologies.

FILIPE _ Yes most of the technologies have this naive motiv. Technologies actually don't have ideologies behind nor religions. Humans make ideologies and religions. But technology and science serve that purpose so to speak. So in some way it's inevitable that technology keeps on progressing also because of universal darwinism as we were talking earlier. We have to make more and more better adapted smartphones if you will, so it keeps on going as if it almost had a soul of some kind... What I mean is that it seems quite inevitable that we'll end up in this situation where you have people that get enhanced and have the right to, whatever they're reasons are, and those who don't. So imagine MU would actually become feasible. Who would benefit from this? The rich probably--- (laughs)

JAMES _ Yes. Well if it's in a destructive process as Ray Kurzweil proposes then I think that it will not be very enriching for any of these people. But I like the idea of taking this technology and making it democratic. Taking .. could make, you know the smartphone

is augmentation. The ear and the eyes are organs of the brain. It might as well be inside of our heads. And this is an intelligent amplifier. And the guy in Ethiopia has one too. And it seems even more amplified in him.

FILIFE _ Yes but you're going to insert that smartphone sooner than the guy in Ethiopia.

JAMES _ Yes. But right now he has access to it and it is an intelligent amplifier. And it is pretty democratically distributed, or evenly distributed.

FILIFE _ Sure. Ok we can go on to the next question which is kind of related to a work in progress of a thought experiment of mine regarding SI and MU.

QUESTION IV

FILIFE _ Would it make sense to iterate human intelligence via mind-uploading to augment it / improve it, and in this manner cope with AI's progress? Imagine the scenario that we would be able to, and at this moment it's not important if the upload has or hasn't personal identity or even consciousness. We'll leave that for now for philosophy. Technically we would achieve WBE to such a level that could turn MU into something also feasible. So imagine I have these uploads and then I would make, I would iterate my own intelligence, since these copies are digital files I could copy them as much as I want and edit them too, just like we do today in genetic engineering. So I could edit the blueprint of my brain and change it a little bit from say filife-copy1 to filife-copy1 million. I would make little changes, alter them, and then take out the copies without consciousness we don't know this now, but I would run these copies in a computer simulation, with an algorithm by then created and would put them to compete with each other and in this manner I could find out that copy nr 4222 billion would be my smartest copy, via my prior editing. I would now select this copy, my most intelligent me, and in this manner I would be more adapted to eventually cope with AI's progress.

JAMES _ That sounds, well A) it sounds feasible and B) there's some speculation. And yes, well you basically described evolutionary algorithms.

FILIFE _ Yes, but there's more to it, because there's something behind this idea which is that

JAMES _ You'd breed the successful ones.

FILIPE _ Yes but then I would take these copies and run them against various scenarios, environmental, economical, biological scenarios like the one we're living today, against viruses, and basically I would stimulate my best possible response against that virus scenario, just because we're living this coronavirus situation now.

JAMES _ I think all that sounds promising. It's a tower of suppositions, you know, that can be successful...

FILIPE _ Yes, that's also the reason why it is a thought experiment...

JAMES _ Yes, it's an interesting thought experiment. It's a good thought experiment. I thought about writing a book called thought experiments, there are so many interesting ones.

FILIPE _ Yeah, sure. Many great ideas made it through thought experiments. The general relativity theory, Einstein, is one of them....

JAMES _ Was that a ...

FILIPE _ Yes it started out as a thought experiment. And just a few years ago, when it made a hundred years, they made observational evidence of the gravitational waves thus proving Einstein's theory right...

JAMES _ Hard Lipson, who's a roboticist at Cornell, created an AI that observed pendulum motion. And it derived some of the basic fundamentals of physics from it.

FILIPE _ I didn't catch it? Sorry? Observing what?

JAMES _ A pendulum.

FILIPE _ Ah ok. A pendulum, yes.

JAMES _ Anyway that's not the main, but he's done some amazing things. He has also done some things in evolutionary algorithms. In a virtual world, I think the solutions to various of these issues, in reference to your thought experiment, if you develop some of these cognitive architectures in the virtual world you avoid some of the safety issues in the real world. Ben Goertzel, for a long time, a lot of people, have been talking about create a virtual world with very very high resolution where you have, you know to navigate in that world you would have to have the same ontology, common sense that you have in the real world, and then raise them through different iterations, better and better

architectures that can navigate that virtual world. And here's the question: How do you introduce ethical education to that creature? But that seems more plausible than raising an AI in isolation and trying to program it with ethics. It seems like if you could raise it in an environment where ethics were kindness and other virtues were rewarded then you would have this hard wired nice thing instead of letting all the commercial drivers evolve it. So I think there's a lot of promise in the idea of developing things in the virtual world. Which is really what you're talking about. You're talking about different brains exploring up in a virtual space and then experimenting and developing them in that virtual space.

FILIPPE _ Ok. Last question. Still in this thought experiment of mine. Could MU prove that basically two different persons sharing an identical blueprint would in fact be identical? Because, imagine all those copies I made, and let's take the original one, without editing. My original blueprint, ok? And I would make the same with your original brain blueprint, ok? And like we do today in genetic sequencing and analysis, you can compare them, both blueprints. In principle of course our genetical blueprints will be different, but not that different, we'll probably share 99% of our genes (or more, but the difference is what basically makes you James Barrat different from me Filipe Luig). But imagine we'd compare our brain blueprints and find out that your blueprint would be 100% identical to mine?! It would be strange, of course, but imagine it as a mathematical hypothesis. Wouldn't this shake our beliefs in identity and personal identity? I'm pretty much ok with being me, myself and I, and if there's a guy claiming he shares my blueprint...

JAMES _ I'll tell you an analogous story. I'm developing a film about suicide.

FILIPPE _ Oh God.

JAMES _ Yeah. Another happy story...

FILIPPE _ Yes you just keep picking them up.

JAMES _ Yes. But it's a critical problem here. And anywhere. But here it's really bad. And they're looking for genetic roots in suicide so they're looking for twin studies. And they don't know which gene, if any, gives you a predilection for suicide, but they do know that there are clusters of genes and they do find that they're pretty much identical in twins. So if one twin kills himself the other one has a very high probability in killing himself even if they were brought up in different environments, even if they were separated at birth...

FILIPE _ That's mindblowing, yeah...

JAMES _ So if that's so and you got two people who are genetically identical and they are not different people really, I mean they are different, but years later they find that they dress similar, they drive the same cars, have married to similar wives and have similar occupations. So I think that blows up our idea of identical personality, in a very similar way to your identical blueprints...

FILIPE _ I had a really great conversation with a geneticist, he wrote a book called "The Psychopath inside us", his name is James Fallon, he's a neuroscientist. And when we were doing the film we were in the locker room testing veterans and burbures for the warrior gene. So we're doing, well it's a long story. Veterans don't get to be aggressive, because when you're aggressive the punishment is death. So if someone's aggressive they wouldn't kill him they would send him off to the desert and vanish. And if he'd manage to survive then fine but nobody could do that. So we found that the veterans and the burbures had very very few instances of the 'warrior gene', few instances of being aggressive. They weren't violent. Even though they're veterans and burbures they wouldn't .. and nobody could remember the last time anybody had murdered someone, not even in their father's memories. So the point in this was James Fallon is a world class neuroscientist, and we were talking about the whole idea of genetic determinism and he basically has this premise that "free will" is an illusion. That the parts of your, and we spent hours talking about this, the parts of your...

FILIPE _ That's also Dennett's position.

JAMES _ Yeah, Dan Denett's brilliant. But Fallon wrote the book "The Psychopath inside" because he was investing into psychopaths and his grandmother told him you should look into your own history because there have been a lot of murderers in our family. And when he did an fMRI scan to his own brain he found that his amygdala was poorly developed and he, himself was probably physiologically a psychopath! (laughs)

FILIPE _ That's ...

JAMES _ Yeah so his mind is blown... He did a ted-talk at some point, if you look him up, a very interesting guy. James Fallon (spells it out). But anyway you know, I got a call in seven (7) minutes...

FILIPE _ Yes of course! I'm sorry, we already...

JAMES _ Keep in touch and let me know how your “*hyperthesis*” is going. Contact me when you’re done. You have to write it and then present it right?

FILIPPE _ I’m already writing it as we speak and processing it. I have to present or deliver it up until Fall this year, i guess September or October...

JAMES _ Then you’ll need some alone time! (laughs)

FILIPPE _ I am! Isolated... (laughs)

JAMES _ This is a good time to get that kind of work done!

FILIPPE _ Yes, it is. You know in the US what they (talking about being barricaded) what they sold the most?

JAMES _ (signals a gun)

FILIPPE _ Exactly! Guns! And in the Netherlands?

JAMES _ Condoms!

FILIPPE _ No that was in Germany! In the Netherlands it was Weed.

JAMES _ Oh, I’m not surprised. I lived in Amsterdam for about five years. Well, they shouldn’t, I mean what they need to do here now is shut the gun shops! Anyway, keep in touch. Are you working in nanotechnology now?

FILIPPE _ No, I’m working in organ transplantation. In molecular biology.

JAMES _ Ok ok. So, great to meet you and great in talking with you, if you have any questions or anything at all just shoot me an email!

FILIPPE _ Sure! Great. Thanks a lot! Bye!

JAMES _ Right. Take care! Bye.

THE END

APPENDIX K INTERVIEW WITH RICK CATTELL

(Computer Scientist, San Francisco, USA)

skype @ 31MAR2020 ~30min + answers sent by e-mail @29MAR



1_ Regarding Whole Brain Emulation and mind-uploading

Q1: Would another me be me? (or just a copy of myself)

If so, then How would you know you are you?

I don't think any emulation of me could be complete enough so that I would *completely* feel like "me," if you're talking about a *robot* me acting in the real world. The biggest problem with emulation would be the loss of my biological body and everything that comes with it. However, it could be an arbitrarily good emulation, such that people would consider it a way to "survive" dying...

If you're talking about my mind uploaded to a *virtual* world that is difficult to distinguish from the real world, then the biggest problem would be emulating the real world in a way that I could not distinguish it from the one I knew. And if I knew it to be a virtual world (as they discover in the movie *Matrix*) it would be a psychological problem, exactly *because* it's not real. Psychologically I want to feel like I'm real, else life seems somewhat meaningless.

Q2: What would represent the biggest risk in achieving something like this?

Ultimately, I think the biggest risk is that we are all living on one small planet, so that a mistake could lead to our destruction as a species:

1. First of all, there will eventually be exhaustion of resources: that is, if people can propagate by self-duplication, we will eventually run out of real-world resources to sustain ourselves.
2. Second, there is a problem with our evolution. Improvements in both individuals and societies have been founded on competition that has ultimately been made possible by the death of existing individuals and societies.

3. Third, we will not avoid the temptation to make improvements to ourselves as we duplicate ourselves, which may backfire if we take a step in the wrong direction... once again because we only have one planet. This could lead to self-annihilation, a super-intelligent dictatorship, or other unpleasant consequences.

I believe it will be a *very* long time before we find a planet as hospitable as Earth, so we're in a "closed system".

2_ Regarding Superintelligence and the Technological Singularity

Q1: If (*When*) it happens (*and we're still humans*) will we even know it (*be aware of*)?

If so, then Could it be because we would be *less bio and more bit* at this specific moment in time?

If not, then..... Would it be because we would have failed to cope with progress?

I don't understand the question. As our intelligence evolves, or if we become one super-intelligence, wouldn't we remember our history as humans? But as I point out above, we could also fail to deal with progress, and destroy ourselves.

Q2: How possible would it be for a *Singleton* of some kind to rule the World?

I think it is quite possible. Again, this is a danger of our "one small world". Over time, the boundaries of countries have become more and more transparent as we have evolved as a species. We are connected with faster and faster transportation, technology, and communication. All of that makes it easier and easier for one entity to rule the World. As an example, in the World Wars a century ago technological progress, military and otherwise, allowed dictators to become "Singletons" of a sort, controlling

much of the world. And now we have another century of technological progress, at an ever-accelerating pace.

3_ Regarding Transhumanism and Posthumanism

Q1: Will *being human* be different in a Post-human Era than it is today?

If so, then Would this mean we would definitely be *more bit than bio* ?

If not, then..... Would this mean that an eventual cyber-sapiens of some kind would in fact be the next natural step in human's evolution?

Yes, and probably yes. In some sense, being human is already different. Doug Engelbart at SRI in the 1960's talked about "intelligence augmentation" through computers that assisted a human in reasoning, collecting, communicating, and remembering information, and his vision has come to be over the decades that followed. Medical advances have also allowed us to augment ourselves physically and mentally in various ways, and genetic manipulation may allow even more. A singularity of sorts will likely occur if mind-uploading becomes possible, and we indeed become "more bit than bio", but that seems a **long** way off to me.

Q2: Do you think politics in a few years will be divided between bioconservatives vs transhumanists?

Yes. We are already politically divided on that. "More bit than bio" is a scary thought.

A side-comment: as humans, we evolve from generation to generation learning from the mistakes of our ancestors, and doing things differently. I wonder how this would change if we lived indefinitely, instead of learning as children from previous generations' mistakes.

4_ Regarding Superintelligence and mind-uploading

Q1: Would it make sense to iterate human intelligence via mind-uploading to augment it / improve it, and in this manner cope with AI's progress?

If so, then..... Could we avoid a singleton of some kind or AI's progress would sooner or later always develop into a *Superintelligence* ?

Interesting question...

- Argument in favor: We have the benefit of starting with the best of billions of variations on human intelligence, and an intelligence evolved from humans might be better at communicating with humans and helping to satisfy human needs.
- Argument against: We can better predict the goals and control the behavior of a designed intelligence "architecture," and we can produce intelligent behavior sooner, indeed we can already demonstrate levels of AI in limited domains.

So in answer to your second question, I'd feel safer with a designed Superintelligence where we know its "control loop" rather than one based on the human brain, whose architecture we do not understand.

But either alternative is scary. Imagine if millions of years ago, apes had a choice of whether to wipe out proto-humans in Africa before they took over the world. It hasn't turned out well for the apes in the ensuing years.

We're smarter than the apes. Perhaps we'd choose a better life for our future and our "children", and perhaps we'd be *less* likely to wipe ourselves out with nuclear war, global warming, or exhaustion of resources if we did that or we were led by some global super-intelligence. But again, scary.

Q2: *Could mind uploading prove that two different persons sharing an identical blueprint would in fact be identical* ? And if so, would this not shake our beliefs in issues such as identity and personality?

If we could completely and accurately mind-upload, yes, we'd end up with two identical individuals, up until their life experiences caused them to diverge.

In some sense, identical twins are the same thing at birth. They sometimes do have some issues with identity, but they seem to handle it all right, in fact many of them like having an “identical friend”. Maybe you should talk to some identical twins on this question!

5_ Regarding Transcending Biology and Human Enhancements

Q1: Is it inevitably going to generate a dystopian society?

If so, then Could/Should we trust our fate in a system run by an AGI?

Q2: Until what extent are enhancements ultimately leading life extension?

Yes, all of this could lead to a dystopian society, and that's scary because we only have one planet. What's even scarier is that we **don't** have “one planet”... that is, if one country or one company does not develop an AGI, another one will. It's like nuclear weapons in the 1950's, or the super-intelligent computer in the movie *Colossus*. Ironically, the development of an AGI that rules the world, if we can design a benevolent one, could be the only way to save us from one that would destroy us!

On the positive side, I see this problem as a fair way in the future, still.

In answer to your second question, yes, mind-uploading would be life extension in a sense, but I think we're reaching the limits in how long our biological bodies can survive.

[These were the answers written to me by Rick Cattell, since he asked me to send the questions before the interview. The actual live interview's transcription follows (2 days later)]:

FILIFE _ So first of all let me thank you in having this interview, and secondly I should ask your permission in recording this, for academic reasons you know, it's not for broadcasting or anything.

RICK _ Yes that would be fine! So let me ask you: Are you a molecular biologist or something? What do you do?

FILIFE _ Exactly. Yes, I'm a biologist and work in organ transplantation, as a molecular biologist...

RICK _ Well the ultimate organ transplantation is uploading the brain... (laughs)

FILIFE _ Exactly, yes... probably or eventually yes, that might be, yes... So after majoring in biology I made a masters and two unfinished PhDs in nanotechnologies and nanosciences...

RICK _ WoW

FILIFE _ Yes, well nevertheless this gave me the opportunity to be part of the development of a couple of biochips, namely nanosensor devices for the detection of viruses, actually... Basically this gives me the scientific and technological background to address these singularity issues, so to speak, hence this PhD now finishing and in regarding these themes, and hence these interviews too...

RICK _ I'm impressed with the breadth of your knowledge!

FILIFE _ Yeah c'mon (laughs). Please... So that being said, and you've already answered the questions by email, I'm looking at them, but I think you might have thought something else about it?

RICK _ Yes, I'm happy to talk some more about it.

FILIFE _ Ok, so regarding WBE and MU, beyond what' you've already said is there something else you might want to add?

RICK _ Well I've been ruminating ever since I wrote this but I haven't got much more to say... You know, I've googled transhumanism and some of the terminology. It's an intriguing topic. It's probably a few decades too soon to have any real impact on us, but it's good to think about the philosophical morality and the possibilities...

FILIPPE _ Yes, but also because you have this psychology background, it seems to me that somehow you make this hypothesis around a psychological continuum after mind uploading, so if there is not a psychological continuum. If we could talk about an upload having consciousness and a personal identity which are quite a few issues around mind uploading...

RICK _ Yes. The closest analogy I could imagine was that with identical twins that start with the exact same memory and hardware at birth but then they diverge from there on...

FILIPPE _ So basically your idea is somehow that it could be something just like me but from there on experiencing makes basically the difference and afterwards it wouldn't be more me but different *me's*... (I'm losing you in the network here... ok you're back)

RICK _ Yes there are a lot of science fiction movies about the topic as well... (laughs)

FILIPPE _ Yeah, not all of them good but yes. And a big sci-fi literature on it too.

RICK _ So then there's the question of, that none of us die, and we just duplicate ourselves, you know do we exhaust all the resources of the planet? Kind of awkward fostering the resources of the planet with eight (8) billion or whatever we got now.

FILIPPE _ Yes but on cyberspace it's somehow different because it would be a file of some kind just as we have today with excel files or acrobat or whatever. It's not exactly that they'd take too much space on the planet as you mentioned, because yes, we're about 8 billion by the time being and in the few years to come and then more and more, so, ok this could be a problem with lots of copies (imagine if I'd make 1000 copies of myself now) but it would be a bit like in *The Matrix* (movie).

RICK _ I guess there's plenty of space in cyberspace, it's not going to run out there. But I meant if we'd duplicate ourselves in the physical world.

FILIPPE _ Ok, yes. So, what else regarding this first question. Ah ok, regarding the second one, you say "I don't understand the question. As our intelligence evolves, or if we become one super-intelligence, wouldn't we remember our history as humans? But as I

point out above, we could also fail to deal with progress, and destroy ourselves.”. Yes, well what I mean exactly is if we take it, well not for granted but, if we *believe* that AI could in fact undergo an intelligence explosion of some kind, in the near future, and so, become actually artificial general intelligence, as general being - human level intelligence, but just a few seconds afterwards this would mean like millions of years of evolution. Well maybe not millions of years for a few seconds, maybe hours or days or whatever, but then it would progress in .. and the real question is would we even be aware of it? It would be so fast, AI overtaking us would happen so fast we wouldn't probably even grasp the concept...

RICK _ Well which *we* are we talking about. The new *we* or the old *we*. The new *we* would, if it were so intelligent it would remember the history... (laughs).

FILIPPE _ Yes the new *we*, in this case would be *we* as a kind of cyber sapiens... So mixed, immersed and merged with technology. That's the new *we*. But the older *we*, that's why I put the borderline here, because before we start being more bit than bio, maybe it won't even happen (the new *we*) because there's a technological singularity of some kind.

RICK _ Yes, so I guess we've covered that question because the new *we* would understand the history and the old *we* would kind of be left behind if it would remain to be still *just human*. So the new *we* would definitely be more bit than bio, yes and the humans left around would be left in the dust I guess.

FILIPPE _ Ok, so you don't want to add anything, or is there anything in particular of these questions you've answered that you'd want to discuss in more detail?

RICK _ Let me see, well the scary thing to all these questions here is that we only have one habitable planet here, so whatever happens to us if it backfires we're in trouble.

FILIPPE _ Well we're in trouble already with the virus, it seems we don't need AI to finish us up.

RICK _ Well we're probably not all going to die from the virus. So, by the way how's it going there in Portugal?

FILIPPE _ Well actually I'm actually in Germany, but yes, things aren't as bad as one could have foreseen, whether in Germany as in Portugal, not as bad as Italy and Spain for

sure, even better than the UK, which are the worse at the moment, I think. But we're all barricaded at home, I guess you too.

RICK _ Yes one state at a time, the US has been doing a lot in that order but I think we're going national now with staying at home. Worst case the corona virus probably wipes out twenty (20) % of us, but here we're talking about something that could wipe out the entire human race!

FILIPPE _ Yes. But you were going to point out something.

RICK _ Yes the scary thing about this is that we're in a small world and that seems to be a theme throughout my answers is that the whole thing of superintelligence is scary because we only get one sandbox to play in. It could backfire. So the good news is that I still think it's a fair way in the future, you know I've been in Artificial Intelligence since 1978 when I did my PhD in AI, and look how little progress we made from then...

FILIPPE _ Yes, but by then AI was going through that winter as they call it, and it got then boosted back in the nineties. And just recently, nowadays, especially with deep machine learning...

RICK _ Yes we're making progress but in a very directed AI, like google being smart in what you asked. But I'm not sure we're anywhere near an AGI, where people are going to, well you know, computers will be like people. Maybe in another thirty years. We'll see.

FILIPPE _ Yes, sure. Talking about Google, you think that it could turn in some kind of singleton? In terms of imagining that it develops so fast until it gets to that so-called AGI, it even now makes lots of our decisions... If I'm driving down the road I'm trusting more in google than in myself. We could get to a point where google would be the singleton itself. I mean it doesn't have to be google, of course it can be another company. Apple, or a chinese company for that matter...

RICK _ Well you're right. Google is very powerful and we use it to make all kinds of decisions so in that sense it's a little scary, especially since the competition of google hasn't done very well. There are a dozen other search engines that just haven't been as successful. But in the sense of google becoming a superintelligence I'm not worried because I know how that software is structured and it's not structured to be super

intelligent, it's just hardcoded to answer questions and it does that. It has a big search index but it's not designed to be an independent intelligence.

FILIPPE _ Yes, not now. But today you have software programming other software, right?

RICK _ Yes but google doesn't design other software. It's really very directed at answering questions to people. It would be kind like turning a car into an airplane, it's a completely different purpose and being an independent intelligence. I guess, what's the closest thing we have as an independent intelligence today is like driving cars? We don't have many things like that. And children's robots that have some independence to walk around and do things.

FILIPPE _ I mean you have lots of narrow AI. Today to detect skin cancer you rely more on AI than you do in physicians. It has a better rate of efficiency in detecting cancer and so on and so on... But I'm talking of AGI. Actually you don't believe in, or rather, you think it's still somewhat distant from now?

RICK _ Yes, I think it's still difficult. And they're not putting that much energy into AGI, they're putting more energy in google directing AI to the goal of answering a question or whatever it might be, but I guess..

FILIPPE _ You know that the CFO of google now is Ray Kurzweil, right?

RICK _ Yes.

FILIPPE _ That could change things...

RICK _ Yeah, maybe... I still think Kurzweil was always more interested in directing AI, you know like his book-reading machine and stuff. I mean, what's an AGI to us? It's very useful to have a car driving itself or to read a book or something but what's the commercial value of AGI or what's the human value to AGI? Because it makes more and more intelligence? I guess it could solve some scientific mysteries for us or something , but even then it would be a directed kind of AI.

FILIPPE _ Yeah, well we're even today totally dependent on narrow AI. I mean a plane doesn't fly without it... Actually 99% of airplane crashes happen due to human error. Not because of AI. But I'm talking of another level of AI.

RICK _ Well it's good to ask these questions. I'm glad you're making these questions. I just think that we're still a few decades away from actually having to deal with the questions.

FILIPE _ (laughs) Ok. So imagine that AI would come up to that level, as many of those guys such as Kurzweil, Ben Goertzel, Eliezer Yudkowsky and many more, so for instance take the scenario as possible, and AI would turn into an AGI and then surpass us. So how could we stop it, you know, from getting to that level? I mean we also wouldn't need to stop it, we could use it for us... I'm just directing it in a Terminator movie kind of scenario, it could turn into a skynet or so. So question IV, is actually a kind of thought experiment of mine where maybe you could avoid a bad singularity outcome of an out of control superintelligence using mind uploading as a technique to iterate human intelligence to cope with AI's progress.

RICK _ Well it's scary.

FILIPE _ I can better describe it to you. Maybe to get you more in context of it. So if we think of MU as a copy of our cognition, well it only would happen as a consequence of whole brain emulation being feasible. So now imagine I'd make a copy or a million of them in just a few seconds (because they are digital copies) and then I would make slight changes between the first one and the last one, just like we do today in genetic engineering, editing. So then I would make some adjustments, by editing and running them in a software with a specific algorithm just like in a game, running those copies against various different scenarios like environmental scenarios, economic, or like today viral pandemic scenarios. I would then select the best response, from copy 500million (for sake of the argument), so this would be my actual response against that environmental scenario. So in this manner I'm iterating my own intelligence, without actually subjecting myself physically to it. And I could come up with my most intelligent me, but I could also predict my behavior against the specific situation explored, and I would have that file, that copy for it.

RICK _ Yes you could do that. That would be sort of fast evolution wouldn't it? Iterating on yourself and in a virtual world so you wouldn't need to worry about exhausting the resources of the planet.

FILIPPE _ Yes because we also do simulations today, only in this case it would be a simulation with my copies. So it could have this predictive character. What do you see against it?

RICK _ Well it's not technologically feasible for the time being and for quite some time, but if you could do that in a virtual world you could create fast evolution. I think I've seen something like that in a movie (being malkovich?). There's a movie for almost everyone with these questions (laughs).

FILIPPE _ Ok the other part of the thought experiment, well it's another version where you just change the variables with the scenarios. So imagine I would compare my blueprint with yours, and they would match 100%. What would this tell you about personal identity. Are we the same person?

RICK _ No, well it could mean we're the same person, but it seems very unlikely one in a billion of a trillion of a trillion chances or so (laughs).

FILIPPE _ Yes but since you can make these gazillions of copies, you would make them from both my copy and yours. The only limitation here is the computer capacity. It's not a biological limitation, if you know what I mean. Because evolution is very slow. It takes millions of years to go from the opposable thumb to the brain to the network and so on... But in this technological progress everything's so exponential, and even with quantum computing coming so strong now, maybe it could be sooner than later, don't you think?

RICK _ Yes well I keep remembering 1978 when we thought artificial intelligence is just around the corner. I still think we're a long way from that. But it's good to think about it. And even if we were identical we would diverge within seconds from experiences...

FILIPPE _ Yes, the psychological continuity, as we already covered that topic right at the beginning. Nevertheless I think it could shake some of our beliefs in what we think of personal identity. Also each and every one of us think they're absolutely unique, right? Other than twins, I think most of the twins actually have this real problem.

RICK _ Yes it could be a good idea to find some identical twins to interview. In some sense when we evolve from parent to child through the years we're making copies of ourselves....

FILIPPE _ Exactly! See, now you're getting to my point, but that's biological. It takes generations to get there. You wouldn't have time to make this kind of experiment. But it's exactly that. We're recombining our information, mixing it, via sexual reproduction, that's what we've been doing for millions of years but you see how slow it is... So we'd compress this recombination of information in just a few seconds or minutes...

RICK _ Yes, excellent point. This is mind blowing... (laughs)

FILIPPE _ (laughs) You're laughing a lot. But I guess you may think this is all still a little bit *sci-fi-ish*?

RICK _ Well, I think it still is in that realm. But I'm glad you are asking these questions. I'm glad someone is asking them, because we are going to come to this eventually, I agree with you on that.

FILIPPE _ Ok. Yes the difference might be I think of it sooner than later...

RICK _ Maybe, yes.

FILIPPE _ You wrote a paper on... "Why is building a brain so difficult?"

RICK _ Yes. And so I'm telling you it's difficult. (laughs again). It might not happen too soon, but on a small scale you could think of it as building a simpler brain and then working a way up to humans.

FILIPPE _ Yes, do you think it possible that you think this way, later than sooner (as opposed to me - sooner than later) because it might have to do with the fact that generally, we humans do have a linear kind of thinking? Because our brains also have a linear kind of evolution, so everything that is exponential is somehow too dazzling for us to fit right in. For instance if you go backwards to the time of the early century, if you would have told people by then what we have now, with computers, internet and smartphones they would laugh at us, and claim we are mad or something.

RICK _ Yes you are right.

FILIPPE _ This has happened in only one hundred (100) years, and we don't need to look further beyond in history because before electricity there's almost nothing technological I mean there are thousands of years with very small little steps, but now if you add the fact that these nanotechnologies and these biotechnologies and artificial intelligence and

cognitive sciences all working synergistically together... So in this view maybe ten years from now is not that dazzling for things to happen... I'm not saying we're going to have MU or something in that time frame, I don't know it might be feasible in the next decades, but I don't bother making futurology and come up with a time frame without the background and being so exponential and everything...

RICK _ Yes of course, we have evidence of that exponential growth, I agree. Well you're going to find wrong that AGI isn't going to happen in this decade, or two. An AGI still might not happen at various levels, you know, we might be able to create an AGI of a worm, and then a bird and then maybe a dog... So create intelligence to what a human could do. Although I think we will have an AGI of some form but the exponential growth you're talking about I still think it's way off before we get that working. But you are right, when we get a certain level where the AGI can work on itself, which is something I think a dog wouldn't be able to do, a dog wouldn't be able to work on its intelligence by reprogramming itself, but it's something a human level AI would, and then we'd probably would see exponential growth. Anyway how are we doing in our questions here?

FILIPE _ Ehhh, I mean we covered it all, also because you had already answered to them in the written response you sent me. So I guess basically that's it. Other than if you would want to add something.

RICK _ Well since I just wrote this yesterday I'm still on it, but we can stay connected. I noticed you are on *linked-in*. So I'll send you an invitation to connect to me!

FILIPE _ Oh sure! Of course, yes.

RICK'S WIFE COMES IN _ Where are you in Germany?

(20minutes of nice non-information based interview kind of chat)

THE END

APPENDIX L INTERVIEW WITH ROBIN HANSON

(Economist/ Professor/ Future for Humanity Institute, Oxford University, UK) [skype@01APR2020](#)



FILIPE _ Hi. So let me thank you for this interview, it's an honour. Secondly let me ask your permission for recording it, just for academic reasons, no broadcasting or anything. And third, just to say a few words about myself, i'm a biologist working in organ transplantation for about twenty years. So after majoring in biology I made a masters and two unfinished PhDs in nanotechnologies and nanosciences. Nevertheless this gave me the opportunity to develop a couple of biochips, actually nanosensor devices for the detection of viruses.

ROBIN _ Great!

FILIPE _ So this gives me the scientific and technological background to address these singularity issues, so to speak, hence these interviews for this PhD, and these questions I have for you.

ROBIN _ Yes. I may have some questions for you later. But let's go on with yours first.

FILIPE _ Ok. I'll go straight to the questions. I read your "Age of Em", of course and a few papers, and I've seen some of your ted-talks, so I'm guessing I already know your answer to this question, but I have to make it anyway.

QUESTION I

FILIPE _ Regarding WBE and MU, would you think another me to be me? Or just a copy of myself?

ROBIN _ I would think it to be me. But I think it's a choice or an attitude or a value decision. There's nothing in the Universe that tells you whether it's you or a copy. It's just whether you accept it as you. And I think that that'll be relatively context dependent in the sense that different people in different cultures and different ways that they were brought up will think differently depending on how it was framed for them in their prior context, so that's part of there being no ultimate real answer that forces everybody to have the same answer. It's how you think

FILIPE _ Sure, depends on the concept we all have on personal identity. Ok, what would you think could be the biggest risk or danger in achieving something like this, like MU.

ROBIN _ Well as always the biggest risk is extermination or extinction. So if somehow it would lead to extinction then that would be a very big risk. It's not obviously connected to extinction but lower level risks are of course any disruptive change tends to upset

existing power relationships and so could lead to conflict and therefore, you know there's the great potential for conflict in this sort of transition, sort of a world without the normal usual things so that's a second, not quite extinction but a large disrupt that people would not agree in the new relationships where who's in charge, or who's more dominant, more richer, more respected, because these sort of changes, change those things.

FILIPPE _ Ok.

QUESTION II

FILIPPE _ Regarding Superintelligence and a possible technological singularity, do you think that when it happens, if it happens, and we're still humans without those big enhancements, would we even be aware of it?

ROBIN _ If there were creatures in the world who were superintelligent, we might distinguish that from a world where there were many creatures of ordinary intelligence who collectively were superintelligent. That is, we as humans for example may be part of a powerful civilization or may be on the periphery looking over and not quite understanding. The more we are part and central of something, the more we will understand the typical kinds of intelligences that are in it, but the more we are on the periphery the less we would be able to judge the difference between a civilization with that contained superintelligences from one of familiar intelligences that were just combined in a spectacular number of ways to produce effective very large bussy? that would be the main way in which we couldn't tell the difference. Obviously superintelligences are powerful, and they have effects on the world, but the effects that super intelligences have on the world are not obviously distinguishable from the effects of many ordinary intelligences organized well.

FILIPPE _ Yeah, well because what I really mean is, AGI could develop into some kind of superintelligence, imagine the intelligence explosion scenario, and a kind of entity. Imagine google would develop into such a level that it would surpass us in just a few seconds by compressing millions of years of biological evolution, and become some kind of singleton.

ROBIN _ Well you're putting a bunch of things into a scenario and of course the more things you put in, the more things I can say about it. Now the way you placed it in the

start is just a vast space of minds. And one particular corner of that vast space and the more you tell me about a future mind, the more I could fill out the scenario in which that future mind exists. Just telling me superintelligence you haven't told me how fast it's growing or if it's artificial or natural or a whole bunch of other things. So I'm just trying to be careful here in addressing the particular thing that you postulated without making a bunch of assumptions about other things, but in your following comment you started to make assumptions about other things and then if you said an AGI could we tell the difference, well there is an obvious focus on the artificial. And then you might say what if our artificial intelligence was as smart as a human, could we notice that and then you're asking about a different scenario because you're making different assumptions. So I just think it's important to go step by step here and not throw in all your loose associations with AGI and you know there are vast possibilities and we have to walk through them methodically.

FILIPE _ Ok. You're right. So let's assume AGI would make that step in surpassing us. A kind of a skynet. The overtake would take only a few seconds.

ROBIN _ Well even if you're walking very slowly it could take a long time if you're small. So the time that takes to pass the line isn't very interesting. What you're presumably talking about is just the overall rate of change of a society that involves some assumptions. But even just the Age of Em which doesn't involve an AGI is a very fast growing society and from our point of view from a person of our world today it would be a very fast world, so you don't need AGI to have a very fast scary world. You could have AGI without a very fast scary world, so again we have to be careful talking about postulating a very fast growing world. If humans were part of a very fast growing future world could they tell the difference between that and a slow growing future world? Yes of course, fast growing worlds would produce a lot of fast changes that would be quite visible and noticeable almost anywhere and yes you would notice the difference. You wouldn't necessarily know the cause of that difference if it was a fast growing world. Could you know if it was AGI versus *Ems*? You'd have to look more closely to see which kind of intelligence was producing that kind of growth.

FILIPE _ Ok, yes. And the kind of resources it would be using too. It would make an impact for sure, yes. Can you still in this scenario envision a kind of singleton?

ROBIN _ The concept of a singleton is just a concept of a strong goal coordination. In principle that could exist now. In principle it could have existed long ago. In principle it could exist in the future. It's not logically connected to AGI or Ems. So I can certainly imagine a future with a singleton but also certainly imagine a world without a singleton. I can imagine that today, I could imagine that today a singleton and I could imagine a world government. We're not that far away from a world government being feasible or even natural.

FILIPE _ Ok.

QUESTION III

Regarding Transhumanism and Posthumanism. Being human would be different in a Post-human Era, I gather and if so could envision us more immersed, more mixed, a kind of cyber sapiens, and so there would be people with enhancements from all the technologies but also those without them...

ROBIN _ Of course. We are quite clearly enhanced humans relative to our ancestors of a few million years ago. If they'd look at us they would see us as somewhat alien or somewhat deviating from their origin and the world they're familiar with. We have compromised in many ways to allow and embrace various kinds of enhancements and changes. There isn't just a binary distinction. There are many gradations along the line from our ancestors millions of years ago to us today. And these kinds of graduations will continue into the future. And whether we will embrace them as us is somewhat context-dependent. So, I'm influenced by this interesting result that I got from a philosophy professor who said that he talks about the teletransport scenario of star trek, and he asks the students here's the transporter...

FILIPE _ Was it Parfit?

ROBIN _ No, he's just a, well more ordinary, but he asked us students, here's the scenario you're about to get in. When you come out will that be you? And after he goes all over the mechanism and how it works he says about fifty:fifty of the students are split in saying if I go into the transporter will I come out as me. Then he asks the other question: You have just come out of the transporter. Is the thing that went in you? And it's a hundred percent. Everybody thinks the thing on the other side that came in before

was them. Now it's the same pair of creatures but when we look backwards and we see the path that leads to us we see that we are willing to embrace all of that like us. When we look forward to things that we might become in the future we are less clearly willing to embrace that as us. So I'm pretty sure that post-humans as you might call them will just call themselves humans. And they will look back and see them as a natural extension of their heritage just like you and I call even if our ancestors might not. So that would be the obvious thing to say about post-humans as they will just think of themselves as humans. There will of course be a wide range of creatures who will embrace more or less the changes that will happen. We see that in our world and call the people who are less embraced to changes backward, primitive. We call them uncivilized. And that's what our descendents would call people like us in their world.

FILIPPE _ Yes. I'm thinking, I don't want to put a real time-frame here (it's kind of irrelevant if it takes ten, twenty or more years), but in the scenario where features like MU could be true, I'm guessing that in that kind of world as there will be those who have access to these kinds of techniques and technologies, so hence enhancements, and those that don't. Just like today because today there are also people who have access to migraine-pills or whatever and those who don't. But then the cleavage is stronger because it generates a dystopian-kind where you have people that are really enhanced and upgraded, more adapted than others...

ROBIN _ I think that it's wrong to state that the future is more dichotomous than the present or the past. I think it's a natural intuition people have based on something called construal level theory which I have written over the years. Construal level theory is a natural way in which we switch between abstract distant things and upclose concrete things, and our habit when we go into an abstract thing is to think that it's more dichotomous, that categories are more distinct, more internally consistent and more distinct across different categories. That's just a more natural way to think when we think of abstract things and when we think about the future we tend to think very abstractly. So, I think it's just natural but wrong to think that the future is more dichotomous than the present or the past. There will be a continuum of enhancements in the future just like there is now. Now, you might have a better case if you say brain emulation that's an all or nothing sort of thing. So then you could say brain emulations would be more distinct from non-brain-emulations in humans and that difference be the difference in the way people talked and reacted to.. Of course I have discussed that particula difference but

thinking in general there's no general reason to think future divisions are more dichotomous or more extreme or more clearly separate than division today.

FILIPPE _ Ok. Nevertheless, and having that said and keeping it in mind, do you think that politics in the future could be divided between bioconservatives against transhumanists instead of a classical left wing - right wing?

ROBIN _ Well it's logically possible, but it's unlikely. The world is a highly diverse thing. There are a great many dimensions on which we differ and politics has to choose one access of difference, out of all the different differences. So that means most of the ways in which we differ are not oriented along a central way that we differ politically. We differ by age, we differ by gender, we differ by wealth. Those are not very strongly correlated with our current main political access of disagreement. So the odds that the difference between being enhanced and not-enhanced would be the main political difference just because it's one of many great differences is not obvious why that should be the one that becomes dominant. I may have a PhD in political theory and I know that the sort of general game theory of politics is to allow into teams that oppose another team. And it's somewhat free what access that is and that there haven't been much political realignments in history where the access changes. So if you look at the main political dimension in 1800 in your thing it was a different political dimension than it is today. And we may be .. meant with politics alignment ten or twenty years ago were better than the alignment in ten or twenty years from now, we'll see, that's one thing people are discussing. So think about the industrial revolution. Industrial revolution was the last enormous change we had on the Earth comparable to the changes that are coming up and the one comparable to that before was the farmer revolution. Now you might postulate that during the industrial revolution most politicians would be industrialists versus anti-industrialists. That's just not true. That was not the main access of political difference during the industrial revolution. Therefore, why would augmented versus not, or computer versus not or more tech versus less tech, why would that be the basis of politics in the future, there are so many possible ways to pick from, why pick that one, it's just not obvious...

FILIPPE _ Yes, well it could have some ground because of historical reasons, political, philosophical even religious, all them mixed together, I could envision...

ROBIN _ Could? Could is very weak. You could imagine the ..of politics as men and as women; You could imagine the domain of politics would be Irvine versus Rule; You could imagine the domain of politics was math versus literature; All these candidate dimensions that we have a lot of emotional residence with that are conflicts or not, nevertheless most of these dimensions that are important to us and we have conflicts with or not, the main dimension of our political..

FILIPE _ Ok, only because also the kind of dystopia it would make would of course be ever dependent on economics and just like I said before even nowadays there are people who don't have access to some kind of enhancements but by then the cleavage would be greater...

ROBIN _ So let me make an even stronger case. It's unlikely through most of history that the main political division is elites versus non-elites, just because that's very unstable. Most political divisions are one kind of elite versus another kind of elite. Because elites tend to have a lot more influence. Whenever you see or say something people say it's not an elite, usually there is another set of elites who are supposedly representing the supposed not-elites. So in fact sometimes people claim that they are the faction for the non-elites. But it's usually a faction of elites who takes their mememntal of declaring support. So similarly the idea that there will be the enhanced versus the non-enhanced, well that's an elite versus a non-elite sort of , so it's just unlikely especially for a main political axis. Just in general they would be struggling with each other to take control of our society. That's common in all societies. There's a struggle over control and the elites have more power to struggle, which is typically factions of elites with other factions of elites to take control. And they may struggle in the name of the non-elites and that may be part of their appeal, but in general they will still be a set of elitess, that of the main actors that are each driving each side of the debate.

FILIPE _ Ok.

QUESTION IV

FILIPE _ This has to do a little bit with a work in progress of a thought experiment of mine, and regarding Superintelligence and Mind Uploading. Imagine the scenario, and I'm making a comparison with nowadays and particularly with me, I work in molecular

biology and I do this regularly, DNA sequencing and genetic engineering, and as you know you can make today your complete DNA sequence. So it's a kind of blueprint, right. And the analogy here is that, and let's imagine WBE does in fact occur it would be feasible, hence also mind uploading, and so I would make a copy of my brain and I'd start from copy #1 until copy #1million (since they are digital copies I can make all the millions I want to) and in between them I would edit these copies and make small changes. Just like I do today in that DNA sequencing I mentioned earlier. Then I would run all these copies in a software, in an algorithm where I subject the copies to a competition in response to different environments or conditions (environmental, economical, or pandemic just as the one we're living now with the coronavirus...). In the end I would basically select my best response, I mean the best copy and so in this manner I could be iterating my intelligence. So once exposed the scenario, would you think it to make sense to iterate human intelligence in this form?

ROBIN _ Well I do discuss that a lot in my book "The Age of Em". So I can just repeat what I just said there. First of all the Em economy would be very competitive and so even without going out of your way you would make different versions of yourself to compete in different parts of the "Em economy". Some versions would succeed more and others would succeed less. And then you would end up making more copies or versions of the ones that succeeded more that would be just a natural result of the economic competition. So even if you don't go out of your way to make some extra special simulation you will naturally have more versions of yourself competing, if you can afford more than one, and therefore the ones that end up being most productive, most appropriate and useful in their relevant context will make more copies. So that's for not just general intelligence but of course lots of specific, you know, which industry are you best suited for, which kind of profession, which job roles, which kind of new augmented technology you would use in which way with which customers, you know you would just be searching in that space of all of those things, and now in addition, when you have young versions of yourself, so you'll get old and no longer be able to be as productive so you'll keep having these younger versions of yourself and so that they can move toward a subjective age of peak productivity after which they would wear out and you will search different ways to train them to be most productive and useful in that world which could include ways of general training but training also specific things and also the process by which you selected various things and in general not only would you vary the training but you'd also just try varying things in the hat? going to make modifications in the emulation

as best you can to understand what works well in the context of the jobs that you were going to give them. And so be part of the process I call those tweaks in the book, that there will be a search in the space of tweaks. And my prediction is initially sort of surface tweaks, tweaks you don't understand very well that you try, and then you will have wins. But those wins will run out quickly, and they will sort of be stuck in the tweaks that you can find and you're probably not going much further until you really understand the brain better. And so there will be this long period where we just don't understand the emulation brains, they're just this spaghetti mess. You could use them to work by then to understand them very well to reorganize. Eventually you will understand the organization well enough to do more systematic reorganization and then you'd end up having more ability to change it. But before that you'd do just much more simpler things like adding and taking away hardware. So the simplest way to modify will be to look at it and say most jobs people need to do, don't need the high resolution of the sight and hearing that the human brain has. has and the human brain is . So you know what? We're just going to take out a lot of that and end up with an Em that is still capable of doing its job and a lot cheaper because it's not going to have all the brain devoted to it. That'll be just one easy win. And we'll also try to take away some areas of the brain with a regular structure and just expand it, you know just copy more of them and see if that gives more memory and more reliable whatever it is people will try simple versions and see if those will work .So the idea is you have the spaghetti but you will try simple variations and some of them will be wins and that will be the early Em where you focus on these simple tweaks that you don't require to understand the brain in detail it just requires that you play with things and see how they work. Honestly that's how most innovation has always gone in history. Most innovation just hasn't come from some deep understanding of things, most innovation comes from just trying variations and that's what we'll do with brain emulations.

FILIPPE _ Ok. Lat question and still in this scenario. Imagine that those, not quite your Ems but that you'd just compare your mind uploaded blueprint with mine and imagine the mind blowing case that they'd be 100% identical. Basically until what extent could we actually claim to be different persons.

ROBIN _ The concept of a brain emulation is the concept of the artificial brain that would behave the same way as you do in the same situations. Now, it may not be feasible to get to that as an ideal. Maybe that anytime you make a brain emulation it's a slight

variation. Now when you have brain emulations of yourself you could make an exact copy of that emulation. That will almost surely have the same behavior as the initial until its experiences make it different then it becomes something with a different behavior. So even if you and I have the exact brain state at this moment then we almost surely would behave exactly the same in the same situation. If you put us in different situations with different and different obstacles then we would become different as a result of experience. And if you put us interacting with each other we will be able to notice those differences. We will notice how we were once the same and how experiences have made us different. And that of course is the same of a brain emulation with its cop. The moment it makes a copy it's very close and predicts what the other would do very reliably and then they start having different experiences and start being more and more different.

FILIPPE _ Ok. Great. These were the questions. I understand you have some too?

ROBIN _ I have some for you. Well, since you're a biologist, maybe you can help me out.

(25min of talk regarding virology and epidemiology because of an *interesting approach* to the corona pandemic suggested by R.Hanson).

APPENDIX M _INTERVIEW WITH DAVID PEARCE

(Philosopher, *Humanity* +, UK) skype@03APR2020



DAVID _ Bare with me. I should have prepared this... How are you doing? Can you hear me ok?

FILIPPE _ Yes I can. You're audio right? Because I can't see you? But you see me right?

DAVID _ Yes, I'm worried I'm going to see myself and it will inevitably distract me from what I'm saying. But I could do it, let's see if this works?

FILIPPE _ Yes! Perfect.

DAVID _ Yes well I don't like to see myself. It kind of disturbs in what I'm saying and I don't evoke the paradise I write about shall we say. (laughs)

FILIPPE _ (laughs). You were writing Hola (in skype). In spanish?!

DAVID _ Well I just was pretending that I'm multilingual but in practice I speak english with a smatch. I sort of have a multi? In my brain, and I cram every foreign phrase I have ever encountered. A mixture of spanish portuguese latin german french and goodness know what else!

FILIPPE _ Ok, because in portuguese it doesn't have the "H". It's just "Olá".

DAVID _ Aha. Ok!

FILIPPE _ So David, can I call you David?

DAVID _ Yes of course call me David, and there's no time constraint or anything.

FILIPPE _ Or should I call you "webmasterdave"? (nickname in skype)

DAVID _ Well I used to be a webmaster, but my web skills of twenty years are out of date and so is the skype of my website which was cutting edge back in 1997 but the web has moved on, shall we say... (laughs)

FILIPPE _ (laughs) Ok. First, thank you a lot for having me and secondly I'm recording this, so I should ask your permission for it, it's for academic reasons not for broadcasting or anything, so other than the Jury I'm guessing nobody's going to watch this, so...

DAVID _ That's fine! Essentially yes, If you want to go for something afterwards or something crops up with something you didn't deal with, it's like I said at the moment

people are dashing around many things I'm more like in playing a videogame than working on my magnum opera or something, so...

FILIPPE _ Ok, sure. Let me just make a quick intro and tell a few words about myself. So I'm a biologist working in organ transplantation for almost twenty years now, so in molecular biology, genetics and stuff. So after majoring in biology I made a masters in nanotechnologies and then went on to do two PhDs, actually two unfinished PhDs in nanotechnologies and nanosciences.

DAVID _ Wow.

FILIPPE _ So nevertheless this gave me the opportunity in working in the development of a couple of biochips, one of them quite successful, a nanosensor device for the detection of viruses.

DAVID _ Good heavens! (laughs)

FILIPPE _ Yeah, but I haven't detected coronavirus yet. (laughs) So, but that's that. So this background enables me to address these singularity issues, such as mind uploading because what really fascinates me is linking bio to non-bio, living to non living stuff. Hence these interviews, and this almost finished PhD, and so yes, having said this I'd suggest we'd go straight to the 4-5 questions.

DAVID _ Ok sure!

QUESTION 1

Regarding Whole Brain Emulation and mind-uploading, would another me be me? (or just a copy of myself)?

DAVID _ First thing is: Is it technically possible? Now, many transhumanists believe it is possible to have whole brain emulation. I'm personally skeptical, because I don't think classical digital computers can solve the binding problem. Now I don't know how familiar you are with the binding problem or if you have read I've written about it but I'm happy to go over it again. What surprises me is that some very very smart people don't as well as do recognise the binding problem as fundamental. Essentially I think what defines our minds is the fact that instead of being 86 billion pixels of experience, which is what naive neuroscientists suggest we should be at most, at least when awake or dreaming we support local binding by individual perception of objects, and also global binding a unitary

self. And yes, I don't think classical digital computers can emulate this even if one were to replace the ones and zeros of classical turing machines notionally with discrete pixels of experience (micro experience). However fast one executes the code and however sophisticated it is and however much one invokes complexity or fast processing power all that happens if you run it faster is that you will have a micro experiential zombie. Now, how binding is possible is a very deep issue and seems to be classically impossible which is one of the reasons why David Chalmers is temptatively a dualist. I'm not a dualist, but I think that phenomenal binding taps into the world's fundamental quantum substrate. This is a controversial view, needless to say. Anders Sandberg, for instance, who has written on whole brain emulation and mind uploading signs crudence just around 5% that in some sense it does involve the world's fundamental quantum substrate. But if one believes, as I do, that phenomenal binding is classical impossible and if one discounts dualism or strong emergence then hasn't got a lot of options.

FILIPE _ Ok, ok. I interviewed both of them too. Sandberg and Chalmers.

DAVID _ Ohh. (laughs). Great! Yes ok.

FILIPE _ So in this case I would say that you'd definitely go with the copy.

DAVID _ Yes, as I said I would think that it is possible to use digital computers for a backup system but I don't think that what we backup is a subject of experience. So I don't even think that it would be a faithful copy of you that you could use in the restoration of you after an accident or something like that. Also there are philosophical issues of the nature of personal identity. Now, I'm personally skeptical in enduring metaphysical egos but I think identity is a convenient myth, but how much this matters is debatable when one wakes up in the morning, one isn't really troubled "Am I the same me person who fell asleep last night?", and I think some people would regard mind uploading as akin to the dilemma one has waking up in the morning but in practice I think it's much more radical than that, because essentially it's creating a zombie. This is a controversial view but as I think other than a totally mainstream view of consciousness or phenomenal binding

FILIPE _ Yes, ok. So still in this question I have a sub-question. What would be the biggest risk in achieving something like this (like MU)? Risk or dangers, to society...

DAVID _ There is one of the rather exotic forms of existential risk. If, on the background assumption that sentient life is valuable, it's something that I would question at least if it's a Darwinian life, but let's assume that life has to be cherished, protected and preserved, then mass destructive mind uploading would in theory be a form of existential risk, inasmuch as one would be creating effectively zombies. Once again I don't regard mass destructive mind uploading as particularly sociologically credible but nonetheless you could say that's a risk.

FILIPE _ Ok. 2nd question.

QUESTION II

FILIPE _ So regarding Superintelligence (artificial general intelligence undergoing an intelligence explosion of some kind), do you think that when it happens, if it happens, assuming it would happen, and we're still humans without those big enhancements, would we even be aware of it? Because it could happen so fast, AGI would develop so fast it would take only a few seconds to surpass us but those seconds actually mean millions of years in natural biological evolution, well not necessarily millions but in being a couple of minutes ahead of us could mean being so much ahead that we wouldn't even understand it.

DAVID _ Yes, this is only the Goerztels, Neary and Bostrom's intelligence explosion scenario. And it's only one conception of superintelligence. As you know there are mainly three scenarios and one is the intelligence explosion which is exactly the one you're sketching; two is the Kurzweilian scenario of a complete fusion of machines with humans that may embrace mind uploading but then there's another conception of superintelligence, probably I'd say it's the minority amongst transhumanists and that essentially posthuman superintelligence will be our descendants that embrace both genetic rewrites, recursively self improving organically robots, but also and profited by neuralink anything that narrow superintelligence zombie AI can do incorporated within you. So you can have narrow superintelligence on a chip. So this crude division of different kinds of superintelligence, given my views on the nature of mind and binding I think there's a sense in which classical digital zombies are invincibly ignorant. They don't understand consciousness, the pleasure and pain access, they don't understand the nature of phenomenal binding, of course in all kinds of primitive domains they can

systematically outperform humans. But nonetheless we can incorporate this kind of intelligence within ourselves. This isn't complete kurzweilian fusion but nonetheless it's a different scenario from the intelligence explosion we were sketching at the beginning.

FILIPPE _ Yes, that one had more to do with the concept of Irvine Good and Alan Turing, of the late sixties. So you'd basically be more inclined to an augmentation kind of intelligence rather than...

DAVID _ Yes essentially it's recursively self-improving organic robots that can encompass within ourselves everything that classical digital computers can do. Incorporated within us, yes.

FILIPPE _ Ok, but let's assume that this kind of intelligence would develop itself, yes recursively but I'm going to make a comparison here with the Terminator's skynet. So let's imagine google or something that could develop that much and turn into a kind of singleton. Can you envisage the possibility of a singleton ruling the world? Or is it too much sci-fi and we shouldn't even bother about it at all?

DAVID _ Though one can in some sense envisage some kind of global brain, I don't think one can envisage global phenomenal binding. And in the absence of a unitary self, I don't think one can envisage a singleton entity.

FILIPPE _ Well, maybe we all would be in that kind of cloud, you know?

DAVID _ Yes, this is it, because right now I can't access to what it's like to be you or yourself, possibly the conjoined hogan sisters who share a ?bridge, it might be feasible in the future to rig up a reversible thumick?..bridge. But nonetheless in the absence of a unitary subject of experience one needs to be cautious. I think talking about "it", in the sense of a singleton. As I said, how it is possible for an allegedly classically pack of neurons to somehow generate a unitary self is extremely controversial. Nonetheless it's tremendously computationally functionally powerful as can be shown in syndromes where phenomenal binding partially breaks down. ? schizophrenia, simultanica mazia, motion blindness and so forth.

FILIPPE _ Ok, third question, and there are two more to go.

QUESTION III

FILIPE _ Regarding Transhumanism and Posthumanism. Would this mean that an eventual cyber-sapiens of some kind would in fact be the next natural step in human's evolution? I mean, with all the enhancements we're getting...

DAVID _ Yes, essentially evolution via natural selection is proverbially blind and based on mutations that are random with respect to the direction of evolution and sexual reproduction is itself a kind of crapshoot. But I think the very nature of the selection pressure changes when intelligence agents can actually choose the genetic makeup of their future offspring in anticipation of the effects of these changes. If you're choosing everything from hedonic set points, hedonic range, genes, allelic combinations, predisposing to extreme longevity, intelligence too, though I don't think there's a singular intelligence gene or anything... The very nature of selection pressure is shortly going to change. How this will play out I don't know. I think generally we have a very impoverished conception of the nature of intelligence. There's a big difference I would say between transhumanists who are and aren't drug naive for example. Yes, it's the next stage in evolution! Unfortunately simply invoking these terms like transhumanism posthuman doesn't say much more than being trans neanderthal or post neanderthal...

FILIPE _ Sure, yes. Ok, still in this question I have a sub-question: Do you think politics in a few years will be divided between bioconservatives vs transhumanists?

DAVID _ Good heavens! (laughs)

FILIPE _ So forget about left wing and right wing, in about ten years or so you'd have bioconservatives in the left wing maybe, and the transhumanists on the right.

DAVID _ (laughs) In practice, you see, that if some truly revolutionary breakthrough in anti-aging in medicine occurs then essentially I think that all the accumulated religious traditional bioconservative objections to the idea of immortality and radical life extension would simply dissolve. Once this *anti-wrinkle cream* outed and was sold out in the stores... So I'd hesitate to say in one sense the distinction between left and right even though horribly simplistic is still going to be relevant for a long time, a long time to come. But in another sense transhumanism cuts across the traditional left-right divide. The very nature of any information based technology makes... yeah my music collection, your music collection is as good as Bill Gates' music collection. And likewise I think the same

is going to be true with virtually anything that can be digital which means most things. So a rather rambling answer here, but in terms of politics we got so many unknown unknowns and obviously something sadly as mundane and horrific as nuclear war is still quite likely this century.

FILIPPE _ Ok, but you already have this political party of transhumanism in the US, so...

DAVID _ Yes, so should one aim for political power or that would be lobbying and encouraging other political parties to steal that cloud. Of course one doesn't have to choose to do both. I mean I'm personally a writer and with difficulty a talker I couldn't cope with the rough and tumble of politics. But yes, heads off to people who are actually engaged in taking political agenda forward.

FILIPPE _ Ok. fourth question.

QUESTION IV

FILIPPE _ So would it make sense to iterate human intelligence with MU in order to cope with AI's progress? So assuming MU would actually be feasible of course, imagine that, and I do that today - DNA sequencing, and I can edit it and so on. So the analogy is that we would have blueprints of our minds or brain capacities, and by that time we would be able to edit them just like today we do in DNA engineering. Then I would make filipe copy 1 and since it is digital I can make millions of filipe2 in just a few seconds. And then I would change, edit, and make little alterations between filipe 1 and filipe 1million and then run these copies in a program with an algorithm to expose these copies to different scenarios. Against those scenarios, and this is a little Darwinian, but my best response, my best copy for that specific scenario would come out, would be selected. This is making sense right?

DAVID _ Yes, given certain assumptions, yes.

FILIPPE _ Ok, so in this manner I'm selecting the best responding intelligence, just like in a game (like chess or go), I would be selecting *my best me*. Also with this copy I could compare my blueprint with yours. And let's assume that they'd match not just 99% but even 100%. What would or could this mean in terms of personal identity? How could two different persons share the exact same brain blueprint and still claim to be different?

DAVID _ Yes, well there's a whole bunch of background assumptions one needs to unpack. The background assumptions aren't unreasonable given Turing machine functionalism but they're precisely what I would challenge in the sense that if you'd want to have exclusively biological superintelligence I would start by cloning Von Neumann, he's buried at Princeton with variations, and essentially hothouse the clones, tweak them, put them all in the program and this recursive cycle of self-improvement. Now, plenty of people would say "Yeah biology is slow and it's primitive, it's carbon chauvinism, we need to escape our primitive substrate". But they have to give an account on how to solve the binding problem the hard problem of consciousness, have to explain the causal efficacy of consciousness, the fact that we're capable in having a discussion over consciousness its nature its varieties its binding etc this would make a lot of scientists uncomfortable because we seem to be venturing into philosophical issues but there are also scientific issues too.

FILIPE _ Hummm

DAVID _ Sorry I haven't really tackled your question but as I said its background assumptions are different from my background assumptions... Someone like Anders Sandberg can probably give you a better answer than I do...

FILIPE _ No it was only different. There are no better answers, only different ones. Ok, last question:

QUESTION V

Regarding transcending biology and human enhancements, is it inevitably going to generate a dystopian society?

DAVID _ This standard argument against anything that smacks utopianism is an historical record of attempts of an ideal society of utopia has almost always ended up in disaster. One of the advantages of upgrading our reward circuitry, this is the focus of my work, in the sense that hedonic greek calibration shooting up hedonic set points, hedonic range, Anders Sandberg here is an example of someone who will say I do have ridiculously high hedonic set points. Imagine a breed of super-anders and so on... If one does create a civilization of sentient beings whose hedonic range hedonic set points has been rat it up to gradience of bliss. And also the very nature of things going wrong

changes. What it means for something to go wrong if there is no suffering. The functional analog of things going wrong can occur but not the gassiness the subjective garsiness. So when one speaks can one speak of a dystopian society in which everyone is blissfully happy? Now if one has other values beyond classical utilitarianism one may point, and highlight features of our society that are sub-optimal. But it really is a civilization in which we have phased out the biological experience below hedonic zero, as will be feasible. We're living in the last century of suffering that is biologically inevitable. Yes, you can't have a dystopia in a traditional sense I would say.

FILIPPE _ Ok. To what extent do you think human enhancements are ultimately actually leading to life extension? Do you align with Aubrey de Grey's version of, you know, we'll live to a thousand years still in this century?

DAVID _ I align with Aubrey De Grey in everything but time scales in that short of revolution. Revolutions are certainly possible. I'm skeptical that older people are actually going to make it, to which I would very much hope that Aubrey's approach can be conjoined with cryonics and cryoethanasia. Because it's going to be cruel telling older people that science is going to find a cure for aging after they're dead. Now of course one can say that science is going to find a cure for aging soon and you're going to benefit but I don't think as yet the evidence warrants this, you know any kind of well controlled clinical trials, prospective trials, they take a long time in humans and sure one can use animal models and so long as essentially not being harmed I'm not opposed to using animal models but they wouldn't necessarily transfer into humans. So I'm completely aligned with Aubrey De Grey...

FILIPPE _ But not on the time scale.

DAVID _ Yes, but as I said, revolutions happen.

FILIPPE _ Sure I could sound a little bit like Kurzweil now if I'd use the argument that it could in some ways be a linear kind of thinking, because exponential progress is very much alive and also all these technologies as opposed to fifty years ago, now they're combining synergistically with each other. So it doesn't seem that much sci-fi that in five or ten years some mind-blowing things could really happen...

DAVID _ Oh, yes this is it. I agree, the unknown of the unknowns and this whole ship of Theseus thing that essentially one can replicate artificially everything but the mind brain

so in that sense one can have quasi-immortality but when it comes to biologically neurons, well essentially it would be possible to regrow or repair them so we come back to issues of personal identity or lack of personal identity.

FILIPPE _ Ok, yes. Great.

DAVID _ Yes, as I said feel free to you know, follow-ups are fine, i can send you some stuff too, because inevitably one tends to gobble enough in the course of an interview. Sorry if my answers may have been frustrating, because I don't know precisely but ...

FILIPPE _ No, not at all.

DAVID _ But it reflects shall we say the diversity of the transhumanist movement. We don't all speak with one voice.

FILIPPE _ Yes, of course I acknowledge that and appreciate it and I thought those were great answers. So thanks a lot, David.

DAVID _ Yeah, good luck with your PhD. I know it can be quite a lonely hard business at times, but you've chosen a quite hot and interesting area and I hope you publish it online rather than let it gather dust in some University.

FILIPPE _ Yes that's my intention, yes.

DAVID _ Fantastic and if you're ever in the UK let me know and we must have an afternoon tea or something stronger.

FILIPPE _ (laughs) Great, yeah, of course. Thanks! Bye.

DAVID _ Ok, take care. Bye bye. **THE END**

APPENDIX N_ INTERVIEW WITH RICHARD LOOSEMORE

(Computer Scientist, New York, USA)

skype @ 21APR2020 ~75min



QUESTION I

Regarding WBE and MU, would you think another me to be me? Or just a copy of myself?

RICHARD _ The problem there is in the question. It seems like the question, you know, would another me, and then you carry on the question and unfortunately everything depends upon the definition of me. In our entire experience the concept of me never involves any possibility of a duplicate except of course in science fiction. As soon as you stretch to situations in which it either gets duplicated or transported, you're already outside the bounds of what the whole concept of me was meant to address. There are obviously a lot of classic scenarios. Somebody gets up onto the bridge in star trek and gets transported down to the ground, and in theory their entire physical body is dissolved in the shift and then rebuilt from scratch down in the ground in the same state it was and then they come back up again and then we assume there's a me through the whole scene and of course there are star trek episodes where something goes wrong, I can't remember how many there are or when they get a duplicate. They get duplicates very often. I think they steered away from that...

FILIPE _ It's just another day at the office for them. They do it all the time.

RICHARD_ Yes. I think they hint at it occasionally, I can't remember whether they do it. But anyway it has been talked about in science fiction a lot. So, just take that one example of an accidental duplicate, they stay behind on the transporter pad and one of them goes down to the surface and then there's a problem because now you got two of them. So you're immediately tempted to say which one is me. Am I me? Because I'm still in the transporter pad or is the other me, me, and I'm just the duplicate. Suppose for sake of the scenario you do get dissolved in the transporter pad and then the transporter accidentally recreates you, so you know you've gone through the process. So there's doubts about the both of you. You weren't left alone. The trouble is when we do ask those questions about which one is me, strictly speaking I think the question can't be asked because there is no concept of genuine versus fake inside the original concept because it never happened. So in a sense they both are you. Really what happened is that you just created another being. So now there are two beings when yesterday there was only one and both happen to have the same starting mental state. And happening to start to have the same mental state is really interesting but it doesn't mean they're bonded to each other and it doesn't mean that one is less genuine than the other so if I

get whole body emulated, if there's a WBE of me, that is complete and full, and if I expire, if I die in the process of creating if I destructively upload myself then it seems easy to say that that new me is the same as me. It's not correct, it's a new me, but on the other hand it's not any different from any other situation where you go to sleep and in the morning the whole system comes back up again and we never ask the question is the me of this morning the same that went to bed last night, because there's a physical continuity. But the physical continuity doesn't matter an awful lot there was an interruption in the consciousness stream and then a resumption and the fact that it's interrupted and resumed every single night (for those who don't have chronic insomnia anyway), is not amazingly different from an interruption that also involves physical movement to another substrate. That is my position on it, and you know I can then go on and address all of the different scenarios in great detail, so just to respond to the one that often comes up what if I'd go through the whole brain emulation and then somebody says I've got to kill you now because we need to keep the new me, that's not the same as me getting dissolved in the process because right now there are two of us and it's still a murder or a suicide if the original now gets destroyed.

FILIPPE _ Yes, well anyway the 2nd me starts its own experience from that time on and beyond that it's another me....

RICHARD _ Obviously it raises all kinds of different questions. If I'm in the process and I know I'm going to get destroyed physically in order to upload, if that involves a moment, even one moment when the original wakes up - it is a murder - to then destroy them. If they don't wake up it's not different from going to sleep and waking up in the morning. It makes the boundary between a normal day and a murder. And I'm sure people can push that razor and split it apart and find all kinds of nasty in-between states. Suppose the old me has a little dream before they're chopped off...

FILIPPE _ A sub-question inside this one Q2: What would represent the biggest risk or danger in achieving something like this, like mind uploading?

RICHARD _ It all depends on the process. They range from you know risks like can you trust the entity? Person or a machine, that is in control of the process? That would be my first question. Is it a group of people that are doing the operation? Is it the AI that is doing the operation of uploading? I'd want to be sure of that because I'd want to make sure that nobody was going to hijack me and carry me, myself off and do something nasty. There's danger in the actual procedure. I'll come back to the question but I got to tell you

an anecdote. I found this anecdote fascinating. I went to have a little operation, take a biopsy of my throat a few years back. Turns out the whole thing was pointless in the end there was nothing wrong. Nevertheless, I went to this one-day operation and had to have anesthesia (the other only time it happened was when I was a kid and had my t...removed), so the anesthesiologist just before I went under explained to me how it would work and one thing he said was this: So we put this thing in your drip we just turn it on and you may go down very quickly or very slowly, but usually, he said, there is a period of time between the time that I turn it on and it will start turning on his effects, and we will walk you from here and take you to the operating room on your own power, we usually don't have to carry you, you can use your own legs, step on to the table, lay down and you're probably will be talking through most of this. At some point, he said, you will fade out but you may be able to ask questions right up to the point you're on the operating table and we'll wait until you go unconscious. But when you come back after you are likely not to remember anything after I turn this tap on. So he said you may be talking to me between here and the operating room but you probably won't remember it afterwards. And just when I was going down I was thinking WoW! So there's a me, there's a track of my timeline that is going to exist! And I'm going to say stuff! And then it's going to be edited out of my timeline! So, I went through the whole thing and when I got up I told the nurse: You've got to tell me did I say anything, because the anesthesiologist said I might talk, and she said Oh yeah you were chatting the whole time! And I thought who did the chatting? Was that me? Because *that me* shut down! (laughs)

FILIPE _ Yes it's a strange feeling that's mindblowing yes.

RICHARD _ It's a lovely little pretaste of whole brain emulation... Anyway there are a whole host of dangers. It does depend upon how you do the emulation because I tend to believe that you actually don't need a neuron to neuron let alone a molecule by molecule reconstruction of the brain. You can get the entire so to speak software emulated very thoroughly down to the level, so accurate that it's down to the level of little bits of mind or noise that you really don't even care about as a personality without having to really map every single neuron. If there was a way of figuring out the functional state and a functional diagram, sort of functional architecture well enough that we know how it works and why all the physical neurons play the role they do you it might turn out that, you know a cluster of a thousand neurons play the role that can be replaced by on little routine. And it's exactly the equivalent that plays the software routine we know why and we're sure of it so therefore the upload doesn't have to involve an absolute copy. But of

course that brings certain questions along with it there may well be very little subtle fine granularity that will be slightly different between the two but there's a possibility of things being miscalibrated or things you know you would never know about. Obviously there are nightmare scenarios where it goes wrong and you end up with a terribly malformed new consciousness just from some kind of error. I suppose that most of the times we just assume that nobody's going to get into the process unless that starts well enough to be reliable. So to answer your question in a sort of summary form there are lots of worries that people might have about the dangers and if I would respond to all of them I'd probably have you pick them out like what do you think about this danger or that danger... Overall if somebody were to build such a process and if I knew how it worked I would trust them.

QUESTION II

Regarding Superintelligence (artificial general intelligence undergoing an intelligence explosion of some kind), do you think that when it happens, if it happens, and we're still humans without those big enhancements, would we even be aware of it?

RICHARD _ So I'm a *believer* in the possibility of an intelligence explosion but there really are two versions of it, ok? And the one that's been talked about the most is, and I don't know if anybody has given it a proper name but, it's the recursive (actually they gave it a name, I keep forgetting), recursive something... There's a recursive bootstrapping where you have one single algorithm which is just satisfying some function that keeps operating on itself and overnight...

FILIPE _ It's actually the idea from Irving Good from the 1960s...

RICHARD _ Yes! And in that case there's no intervention by the engineers and they just pushed the bottom on the right algorithm and it explodes and it becomes super intelligent by itself all the way up. And then there's the version number two, in which people construct a superintelligence. They do all the work needed to get it up to near the human-like intelligence and then they duplicate it massively and get all those duplicates to help them, so in other words, they amplify their own capacity to develop the next version and then they all, the humans and these intelligences, which are not super yet, they are all work together to design the next generation...

FILIPPE _ An augmentation...

RICHARD _ Well it would be kind of an augmentation but not an augmentation of each and individual human brain but just an augmentation on a scale, you know you would just speed up the scale, speed up the work. In the paper that I wrote in the book (Singularity Hypothesis), with Ben Goertzel, the thing that we discussed most was the idea of number 1- amplifying by multiplying the quantity of computing power as I just described so you have many duplicates; And also speed ups, just pure electronic speed ups where each individual intelligence in that set is able to have its so to speak clock speed increased enormously so that they are all perhaps only as intelligent as we are, but they're working at maybe ten, a hundred or a thousand times our clock speed. Those two factors combined are likely to lead to a situation where, as soon as we get relatively easy to build you know say the size of a laptop pc, you know something that can be easily reproduced, as soon as we get an intelligence in such a situation, it doesn't take a whole city block but it's just small to build it .. and we can build lots of them assuming that we can control them to all work together towards that task of building a new version of themselves. That will speed up but it won't be an overnight process but I still would call that an explosion because it's likely to go from the first human level intelligence of day one, and then within about a few years (it could be very short, it could be a few months), but probably a few years to go from that to the same size of intelligence with either a thousand times the speed of human intelligence or in some other sense a thousand times intelligence. That's a different concept that isn't well defined. I still would call that an intelligence explosion but it is very different from the other version that I first mentioned. And the first one is the only one to which you could answer Yes to your question, in other words, could it happen without us knowing it? Yes in that first case. In the second case you would notice it pretty very much because there are a lot of people working on it and it's not going to happen overnight, it takes a lot of work by a lot of people, human and intelligent machines to make it happen. And if you were to ask me about that first situation - I am deeply sceptical, about that first scenario even though a very large majority of the people who talk about the singularity just seem to automatically assume that the first scenario is not just possible but likely. I think their reasoning is terrible and my paper that I wrote for the trip symposium in 2014 gave all the reasoning why their reasons are just terrible just awful.

FILIPPE _ Another quick sub question inside this one: still in the realm of superintelligence and the singularity. How possible would it be for a singleton of some kind to rule the

world? In that first improbable one so to speak scenario you talked about, but if it might occur we'd eventually end up with some sort of supersmart google ruling us. Can you envisage this?

RICHARD _ I can envisage that under a number of different circumstances but again most of those things don't map on to what other people envisage so it's my interpretation of what that might be, that is not the same as theirs... So just to be clear were you talking about the first kind of explosion, right?

FILIPE _ Yes, because if we were to be aware of it we'd have sufficient time in stopping a bad kind of outcome where this superintelligence would actually make decisions for us. It doesn't necessarily mean it would have been bad decisions, I mean we already got them in a human scenario... (laughs)

RICHARD _ Right. Let me just answer my wife there for a second. She's trying to do an operation...

FILIPE _ Sure yes of course!

RICHARD _ She doesn't know that I'm talking here. Let me just catch up here. Alright. There's no crisis and she knows I'm on the phone. So, if I take the question in two parts you know according to which type of explosion happened. If it's the overnight explosion recursive intelligence, the circumstances under which such a thing would arise would probably involve a very simple motivation system and, so if something arose through the night then almost inevitably it becomes a singleton. But you know, the trouble is, in a sense all bets are off because, all bets are off meaning it's impossible to determine because if a superintelligence arises because of an overnight recursive explosion then the type of intelligence, the end product is going to be the result of a complex non-deterministic process, we don't know what its motivations would be at the end of the rebuilding, we don't know if it preserves its motivation function if it changes it, or if it gets really to that state and says "I would really like to have some friends", and so therefore "I will create a thousand of me, and we're all going to do stuff together", we don't know if it would go around looking for other rivals and just kill them off, so I don't think that, you know a lot of people claim, or talk as if they would know what such a thing would do, Steve Omahandru obviously and many others, and I think they're all wrong, just because you don't take a recursive bootstrap and then it can develop so fabulously that it can go

from a situation where it is really pretty dumb and then all the way up to become something which has an incredibly powerful intelligence. And they simply assume that even though you don't know how the intelligence works you sure as heck know what the motivation of that thing would be. So you can think of it in two things mostly: intelligence and motivation. And what they're saying is it's going to have, you know overnight, from last night to this morning, it acquired an intelligence so far beyond ours that we can't even guess what it would be like. What it could do, what its power would be, how it would work. So they admit that, the people who go through this reasoning, and then at the same time they start talking about the other half of that intelligence which is the driver. The controlling mechanism, the motivation. The goal system. They talk about that as if it never changed, it stayed through the whole event, nothing ever happened, and we know it will be like this, and we can do little mathematical theorems and say it'll do this and Steve Omahandru and others will say it will inevitably become psychopathic. Now the fact that they can admit that half of it is completely unknowable without missing the beat, say the other half of it, they absolutely know how that would inevitably go. You know it has no choice, it would definitely do this or that. That strikes me as, I'm sorry but it's an embarrassingly silly argument. I don't think it can be upheld in any way you know, and you could go deeper into it and attack it on lots of other grounds but for that reason alone to be sure about one half of it and not sure about the other half just doesn't make any sense to me. So, my answer to that is who knows? I still don't believe it'll happen for lots of other reasons but if I'm told to assume that it'll happen I just would say that there's no way of knowing if it ends up like a singleton, or if it makes lots of other buddies or knowing if it's friendly or not.

FILIPPE _ You remember a few years ago, I'm not sure if it was a google chat-bot or a microsoft chat bot...

RICHARD _ Oh yeah! Microsoft yes!

FILIPPE _ They had to shut it down after just 24 hours because it got really interested in things like Hitler and was obsessed with taking power and they thought "wait this is taking a strange way, we've got to stop this", and we're still talking about narrow AI. Not general AI, so imagine what could be....

RICHARD _ Yes. The interesting thing about that is that it didn't really develop. You know a lot of people drew a lot of conclusions from that but it wasn't even barely an AI. It wasn't even insect level intelligence. All it was doing was, effectively scraping the internet,

distilling it all and then pouring it all out to a distilled average of everything you find on the internet, which is just *human stuff* and who would have guessed that if you'd distill all the internet you would find it dominated by ludicrous conspiracy theories and all the rest of it is just what everybody would have guessed if you'd distill the internet. So distillation, of some total of all the chat on the internet is a very different thing and you know although it was an interesting episode I wouldn't interpret anything deeply about it. I think it was more a comedy than anything else, but I take your general point which is yeah even that much was unexpected! They didn't really expect that! They should have done, but they didn't and it does speak to the fact that there's a lot of unexpected stuff that could happen yeah! So the other part of your question which was about the singleton, just going through it briefly without a lot of detail, if the type of AI that is built is done, you know, is still an explosion but in the course of a few years and is done by people in the manner that I just described there is an immense amount of control and distributive decision making about how it ends up and just to summarize all of that is a book length, you know topics obviously, but to summarize it, all my feeling is that people would decide that actually they need a singleton or a cluster, you know a singleton cluster a redundant cluster and they need it really because once the general technology is available the one thing you don't want is for people to go off into a dark corner and copy it and have the freedom to go and build some crazy version of something that was built very carefully by a very large group of people the first time around. So you probably will get to a point where that first carefully builds explosive AI that takes a couple of years where you want to set it up as a singleton, in other words you say: ok you now have the authority to make sure that it can stop the dangers of AI from then on. Because there are dangers once you get beyond that stage and I suspect everybody will agree all, all of the smart people will agree what happens with the rest of humanity I don't know. Will they agree to create a singleton or a redundant singleton cluster.

QUESTION III

Regarding Transhumanism and Posthumanism, assuming the fact that being human will be different in a Post-human Era than it is today then would this mean we would definitely be *more bit than bio*? a kind of cyber sapiens, so not that biological anymore...? Is that the natural occurring path of our evolution?

RICHARD _ We would definitely not simply remain just as bio. Bio is too fragile. Whether we end up as some kind of hybrid of some kind of a positronic brain inside a biological body, you know a hardened positronic brain that can't ever get damaged but nevertheless inside something that we'd prefer to keep as biology... That's a question I can't easily answer. I suspect that the, I think the most important answer to that question is I think we'll have the choice. I think a lot of people will immediately want to go and explore, you know to use their choice and to go and explore every imaginable variation, a lot of people would just go straight to silicon or whatever is the replacement for silicon, or optoelectronic, well just call it positronic brain, it's going to be something that, some unknown technology that is going to be better than silicon. And a lot of people would jump into that with both feet just because it's going to be so wacky and so fun. But I think also people would be able to pour themselves back and forth between different substrates. So they will probably have choice and they probably won't have just one choice, they'll be able to go back and forth. Some people will say I just want to stay with *the biology* and some people may choose never to leave *the biology* and I think that what will happen is everybody will respect that. If you don't want to, I think it's sad, but you can stay in biology and some people may find hybrids are comfortable for them, like having the positronic brain in the physical body, they don't look any different they don't behave any different, they just have a deeper way to think and think faster or they may indeed just say I don't want to think faster, I want to think the same speed I do now, but I kind of like having the positronic brain because I know it's robust and if anything goes wrong, this will survive and I can just get a new one. So I think there'll be an amazing variety of choices and different people would choose all of them. Some people may choose all of them at different times. But I think the essence is, the crucial true things are there will be choice and there will be the possibility in making changes in that choice. You will be able to port yourself from one substrate to another and the varieties of that choice will probably be huge and we could guess them now or lay out some of the possibilities, but frankly I think all of the possibilities you could imagine will probably be available.

FILIPE _ Ok. Another sub-question and because of those enhancements we expect to be bigger as time and progress goes by. Do you think it is possible that politics would be in a few years divided between bioconservatives on one side and transhumanists on the other? Instead of the classic left wing and right wing?

RICHARD _ I think I'll answer it with a question: If we go back to, when I was kid listening to music in the sixties and seventies, well let's go back to the sixties, I was a small kid then, and there were only a kind of only few big musical choices. You were either into kind of old school forties and fifties kind of easy pop, or you were into new rock & roll...

FILIPE _ progressive rock by guys like Led Zeppelin and stuff...

RICHARD _ No I'm talking about before. No I'm talking about the fifties dance parties and there were just a few small categories and you were either the kind of person, say in the middle of the sixties imagine you're twenty years old in the middle of the sixties, you and I would be into the new swinging rock & roll kind of stuff or you would be an older person and would dismiss all that stuff and would say "No I prefer these old singers and dance and I don't like all this modern stuff", and obviously there were more choices that some people liked but you could probably count the major categories like this and in these categories there was one big divide between new style rock & roll and old style not rock & roll. The one big divide, which is a sort of like the left-wing right wing. So the people on the right liked the music of the fifties and the people on the left liked the new loud brash rock roll that would break all the rules. Now, today, what do we have? There's no pop charts anymore. There are, but they're heavily managed. Everybody listens to something totally different and the music scene is fractured into...

FILIPE _ A million pieces...

RICHARD _ Yeah a million pieces but I was going to say I would guess you could name fifty (50) *genres* now probably that people really exclusively are interested today and they don't have much interest outside their little exclusive *genre*. At least among people who really care about the music. Some people just don't care, drift through. And I think that is representative of what will happen in the future as far as politics goes. So the division between the bioconservatives and not will actually be a big spectrum. You will be able to name 50 different sets of opinions about that. But the whole question about how politics itself, you know decision making will play out, is going to get kind of fuzzy because I suspect that society will break into domains that want to live a certain way and they will basically appeal to the AI, people will probably like the fact that they can delegate a lot of stuff to the AIs and say it's hard to manage it's just too complicated and so the economy, as soon as AIs show signs of saying look you can walk if you want but you don't have to I'll give you the food you can have a reasonable amount of luxuries don't ask for too much but you can have a reasonable amount, you can go for a vacation and

you can do this, you can't do that but .. and pretty soon you'll have a situation where a lot of people kind of appeal to the AIs to say I want to live a life where people do certain things, can I have that? And the AI says yeah you can have it up to a certain limit and give them what they want and people will fracture into certain groups that want to live certain ways but one thing they'll find out is that the trade off for having that ability to live the kind of life they want with the kind of people they want will come with a trade off that they won't be able to go to and impose it on anybody else. And I think they will live with that trade off and as a result politics will become very different because there won't be bioconservatives arguing .. For example there will be people still, for a long time who say that you can't abort a fetus after the sperm meets the egg. And they will be very loud and they will be very angry and they will continue to be very angry probably for a long time. But they just simply won't have any power. And if somebody else says that's not my choice I want to have a different choice. Those first people won't be able to maneuver, they won't have enough control over the system to impose their will on anyone. So the politics will fracture, so in other words it's complicated to answer your question but yes they will be different they will be split. I think that some people will stay bioconservative for a long time and then some of them will allow themselves to expire and allow themselves to die, because they don't care or they don't want to. I think many of them will have a last minute conversion and say "actually I'm a bioconservative but now I got cancer (or whatever) and I actually would like to have my body changed and have the genetics changed in such a way that I don't have this, and they will appeal to the AIs and all of the sudden they will become bioconservative but not quite. So I think that is roughly how it will go. Some people will jump rapidly and they will be walking around as robots and will make themselves into metal robots instantly because they want to. Other people will stay *in the biology* and then there will probably be a spectrum in between them along with lots of other choices we could go into. There's a whole book in each one of your questions!(laughs)

FILIPPE _ We're almost at the end.

QUESTION IV

Regarding Superintelligence and mind-uploading. Would it make sense to iterate human intelligence via mind-uploading to augment it / improve it, and in this manner cope with AI's progress? So I have this kind of work in progress of a thought experiment. Basically,

and today I work in molecular biology in organ transplantation, so I do a lot of DNA sequencing. So we compare a whole DNA with another one and so imagine the scenario we would do something like this but with the blueprints, the uploads. I would then take my upload and just as I do today in DNA manipulation I would edit it, and because it's digital I'd do millions and billions of copies, they would all be the same. And now imagine I would put my edited copies, slightly changed copies, compete with themselves against a specific scenario that could, for instance, be a viral outbreak or simply playing chess. So I would then select my best me in playing chess and in this manner iterate human intelligence.

RICHARD _ Yes I can see the scenario. There are questions. You can make new versions. Do you incarnate them? Or do you keep them... Do you give them bodies?

FILIPE _ No. They are just *uploads* in cyberspace. Uploads of my brain capacities in the cyber realm. Billions of them. With small differences by editing them and then with a plot, you know by graphing this I would select the best ones in a darwinian kind of way I would get the best chess player or against a scenario like in a pandemic I would (Clockwall bangs)...

RICHARD _ Yes.

FILIPE _ Sorry I can't hear it'll take just a few seconds (shows clock on the wall).

RICHARD _ Oh beautiful..How old is it?

FILIPE _ I don't know maybe like a century? I think it's late twenties...

RICHARD _ Cool. Yes, if you get a large number of duplicates then they're all going to be at your adult level, because you copied your adult self, and if you try making changes to them we'll assume for the moment that you understand the structure and how it works. So you don't make random changes that are likely to be insane or badly malformed, so in other words, for example you understand where the pain system is, inside each one so one thing that you won't do is go and change that. Some of them will

FILIPE _ Yes just like in DNA technology I know which ones to take out and which don't.

RICHARD _ Yes, assuming that you have good knowledge and the changes that you make don't create ethically bad copies. You'd have different intelligences, you know different ways of thinking about stuff. The next question is what would be the darwinian

situation that they would be subjected to? Because there are different possibilities. You could...

FILIPPE _ Well I would run them in a real kind of crisis like climate change kind of scenario (forget about chess). So like in a game you have this algorithm of competition and I would subject lots of, yes we could call them avatars...

RICHARD _ Yes let's jump forward a little bit because this is a similar situation I thought about. So if you are interested in testing lots of different kinds of intelligence mechanisms, then whether they derive first of all from an emulation download of your brain or from an AI design that you have, that never came from a human being but just involves that either mechanism. If you're involved in just wanting to do lots of darwinian testing on different variations on the design you can either build a complete system, so in the case of the brain emulation of you, a complete version of you that wakes up in an artificial reality and it finds itself in a laboratory, reading papers about a virus outbreak, and it knows that it feels like it wants to solve this virus outbreak problem, and doesn't think about going home to get supper (laughs), because it's just sitting there in the laboratory and the only times it ever gets woken up it feels well fed, and wakes up and sees a virtual laboratory and thinks and tries to work on the problem and produce its ideas, but it is nevertheless a whole version of you, that wakes up and thinks "my name is filipe and so on", and it's aware of everything because it can think about its past history because it's a complete version of you but it just happens that somehow we make it so, you know, that it's just concentrated in working in the problem of the virus outbreak. And then you do that a million times, a million simulations all in their own separate virtual laboratory each of them tries to work out the outbreak scenario and each does it in a different way and you select the best solutions. Now, that's one way of doing it, we'll call that *the full system test* but you could also do it in a different way. If I understand which part of the brain is doing the actual thinking, the core thinking part of it, we could just take that part out and it has no longer the rest, so there is no longer a consciousness. So this is a unit test. This is a subsystem test and you don't have the danger of building a consciousness. But you only test the thinking parts of it in a very tight environment where you don't give it a lab to look at, you put in patterns and it relaxes and comes out with these solutions in patterns but it doesn't think about itself, it has no self to think about because that has been all taken away. So the first of these is going to take a world of ethical questions, a gigantic raft of ethical questions and also resource questions because it involves a greater use of physical resources and I suspect that a simple

answer to that question is nobody will actually try to do that first full body system scenario, because if they understand enough about the way the brain works and I think they'll have to, to get the whole brain emulation thing to work in the first place, if they understand enough they would do subsystem tests on the mechanisms inside and until the extent that they use your mechanisms, your biologically developed intelligence mechanism, it will look like you, it will think in the same way as you but it won't be you, they'll make sure that it has not the life memories that you have, personal life memories, background knowledge perhaps, but not episodic life memories (in psychology terms). So I think that is feasible and you could try to do it, especially in emergency situations, like the virus crisis. But you know, that kind of thing will also be a key part of a superintelligence explosion of the type 2, which takes a few years. That would be an absolutely core part of it. You've described exactly what will be going on most of the time during those couple of years of development. Much repetitions, many variations, experimentations, will be on the core mechanisms of intelligence, not mostly the whole system so the whole systems tests mostly not.

FILIPE _ Last question! Could mind uploading prove that two different persons sharing an identical blueprint would in fact be identical ? And if so, would this not shake our beliefs in issues such as identity and personality? I have to pose the mathematical possibility / probability in matching a 100% identical blueprint.

RICHARD _ So there are two versions of your question and you probably just mean one of them. So if we were looking at the structural design of my brain and your brain, and it turns out surprisingly that every mechanism in here, does exactly the same. It turns out that there's this very close correspondence between most people, and it turns out that you and I just happen to have exactly the same design...

FILIPE _ Yes, well in the previous experiment we were talking about imagine that my copy would compete against a specific scenario, yours too, and both of ours would respond in the exact same way!

RICHARD _ Right. But they'd be different in the accumulated experience and knowledge. Because they had different upbringings. Now, here's a question: How much of a personality is the underline design and how much is the experience and knowledge? And we all know that the real answer is almost all of it is in the experience and knowledge. Identical twins might be biologically absolutely the same but one thing we don't know of course by the way is how much the actual mechanisms of intelligence are not genetically

determined but are temporable during development. So there's no clear separation between mechanism and content, actually the content can change and influence the mechanism, and you can build new mechanisms in the same way that we can upload new software onto one hardware computer is obviously very flexible so there might be some flexibility and people might develop in thinking mechanisms that really are the mechanisms by which they think and it's not in the DNA but in interactions between the DNA and the mechanisms the DNA design of that brain and its interaction with the world and as a result of the interaction of the world one person has a way of thinking about a certain problem that is so deeply embedded that is actually a mechanism something they use all the time and so deeply down there that they don't even know it's there. So I suspect what that means of course is that it makes your scenario less likely of course, but we're still entertaining the mathematical possibility that even so, it just so happens that two people have the same mechanisms and the only difference is in the actual life experience and the content and the knowledge. I think it would mean very little, I think the two of them would think in much the same way about any problem but that would only be the case as long as their past experience was not part of the solving problem, right, so for example I was going to choose chess, but that's lousy because each of them would have looked at different chess games through their entire career. So I think that even if somebody found that there was a circumstantial identity between two brains there wouldn't be any question of being the same, they would just happen to think in the very same line, and they'd probably have a good time laughing if they ever got together because they would be wanting to say the same thing to the same situation they'd keep laughing at each other because they would be doing the same kind of thing, but nobody would even suggest they'd be even remotely the same person. They'd be far more different than a mind uploading duplicate. Because that would start off with two beings. If you uploaded two copies of me, they would have a dispute about who's the real me and then there would be questions about who's the real me, and I already said it becomes a meaningless question to me, there just happen to be two of them. I think it would be funny if .. I suspect that we'll eventually find that people are so incredibly different. But there's not even a remote chance that... Every single person on the planet has a different way of thinking about the same problem if you looked carefully enough. People are amazingly different inside...

FILIPE _ Exactly the case you were talking about the identical twins, and I saw in a documentary or something about identical twins that don't even know of each other, you

know, they have no relation. One lives in the US and the other in Asia or Russia and by the time they are in their 30s they get to know each other. They drink the same beer, smoke the same cigarettes brand, their wives are very similar and stuff like that. Yes they laugh a lot of course but they also feel a little bit you know dazzled by it. It's frightening in some way to feel this way that there's actually another guy who looks like me, thinks like I do and feels like I do...

RICHARD _ Yes. I think I never really watched a documentary... I'm fascinated by those I find really intriguing. I do worry a little that sometimes the storytellers, the documentaries and people that write the books do want to kind of chazz it up a little bit because it makes such a better story. So I never met such people but I'd love to meet them in the flesh. I can well believe that there are these things and some of our behaviours are amazingly determined. I think that if the genetics is absolutely the same I would not be surprised if there are amazing similarities. I think I would be surprised if it went so deep you couldn't find some dramatic differences. Obviously if one of them gets raised in China I don't know if there are any like that one speaks chinese and never spoke english and the other never spoke chinese I would find that really quite interesting. Because whether or not the similarities are sustained across a massive language barrier, that's fascinating. But yes, I take your point. I don't ever underestimate the influence of genetics on behaviour. Not just because of people but I'm a crazy cat guy. We're in a tiny little rural neighborhood of five or six houses and it had the reputation over the years where everybody said you can drop your unwanted cats there and the people have been looking after them. And it's been so for twenty years, we don't know who's doing it but we do know that somebody admitted recently that the near-by folks spread the story by word of mouth that here's the place to drop them off and it has been happening since before I moved here. So we have found homes for nearly thirty (30) cats over the years and many of them we've taken in ourselves and I still have some in the house. I've got to know these animals and socialize a lot with them and I have got acutely aware of cat psychology because I take these wild cats and I got them to the point where they are completely unsocialized and they start off absolutely terrified of humans and now they want nothing more than to interact and come up close. And one of the fascinating things is to observe genetic similarities. So there's one pair of kittens that we know came from her mother and they were separated from their mother in a storm, but survived together and I took them in when they had grown about a year and a half later when they became adults. Then the mother stayed outside and the two daughters have been together in the

barn the last eight to ten years and the mother was outside the whole time and I fed her everyday and a year and a half ago I brought her inside as well. The mother does not know, or care, that these were here kittens, I'm pretty sure, but they have psychological quirks that are absolutely identical. They are so similar that it makes me laugh all the time, they have an approach characteristic of going back and forth, they both have it and they grew apart, there was no social interaction.. Pure genetics. Absolutely fascinating.

FILIPPE _ Yeah genetics is very strong yes. Ok Richard, thanks a lot! This was a great interview...

RICHARD _ Yes thank you for the chat. I enjoyed it a lot, it's great to talk about these things because I keep out of the community these days. I do a lot of other stuff so...

FILIPPE _ Sure! I hope things go well with your father in-law. Let's expect that, and yes once again thank.

RICHARD _ Yes, so good luck with your thesis and stay in touch!

FILIPPE _ Sure! Ok I'll be seeing you! Thanks a lot!

RICHARD _ Ok! Bye now!

THE END

APPENDIX O _ Calculus for Plot A & Plot B _ WBE feasibility plot (Part II – WBE)

Table 19 – Calculus and critical values attributed for Plots A and B (WBE Feasibility).

RELEVANCE	connectomics	Brain Scanning	Computing Power	Non-Organicism	AGI
funding	YES	YES	YES but	undefined	YES but
autonomy	YES but	YES but	YES	undefined	YES but
NBIC	YES	YES but	YES	YES but	YES but
hype	YES but	YES but	YES	undefined	YES but
experts	YES	YES	YES	YES but	YES but
TOTAL	8	7	9	2	5
RELIABILITY	connectomics	Brain Scanning	Computing Power	Non-Organicism	AGI
funding	YES	YES	YES	NO	no
autonomy	YES but	NO	YES	NO	no
NBIC	YES	undefined	YES	no	no
hype	YES but	undefined	YES	no	no
experts	YES	NO	YES but	NO	no
TOTAL	8	-2	9	-8	-5

TO ACHIEVE WBE	description	numerical values	
YES	critical	2	
YES but	not critical	1	
undefined	irrelevant /unknown /not applicable/ does not affect	0	
no	negative effect	-1	
NO	definitely negative	-2	

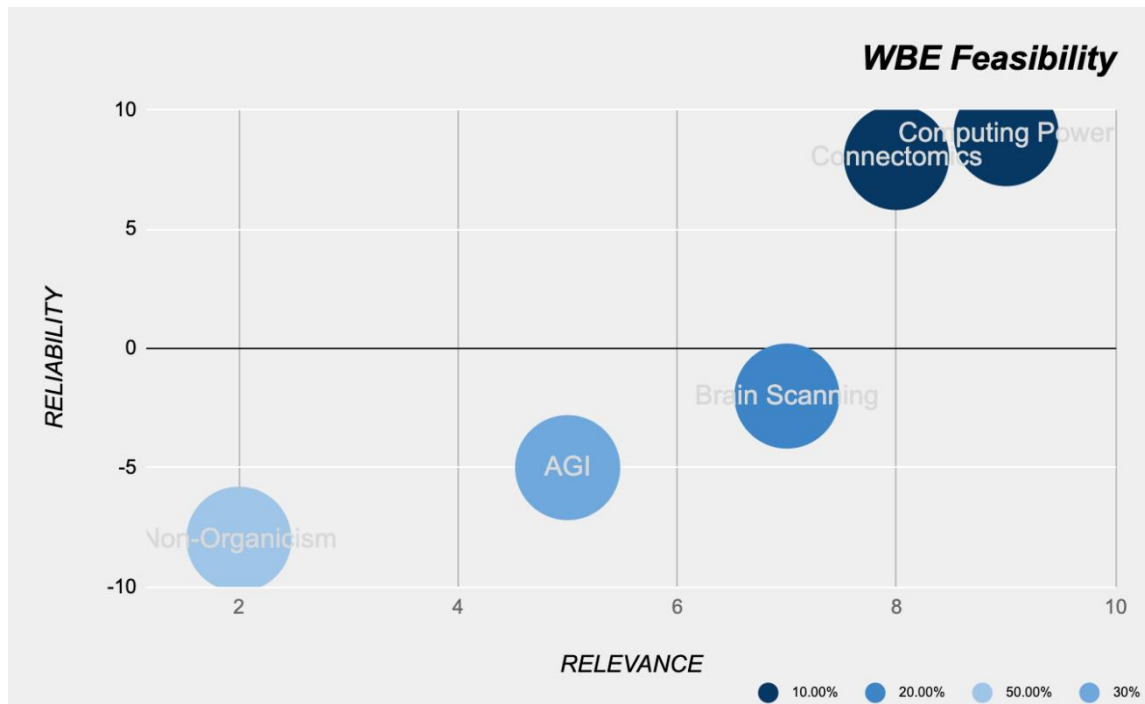


Figure 06 – (Plot A: WBE feasibility regarding critical issues. (Luig, 2020)

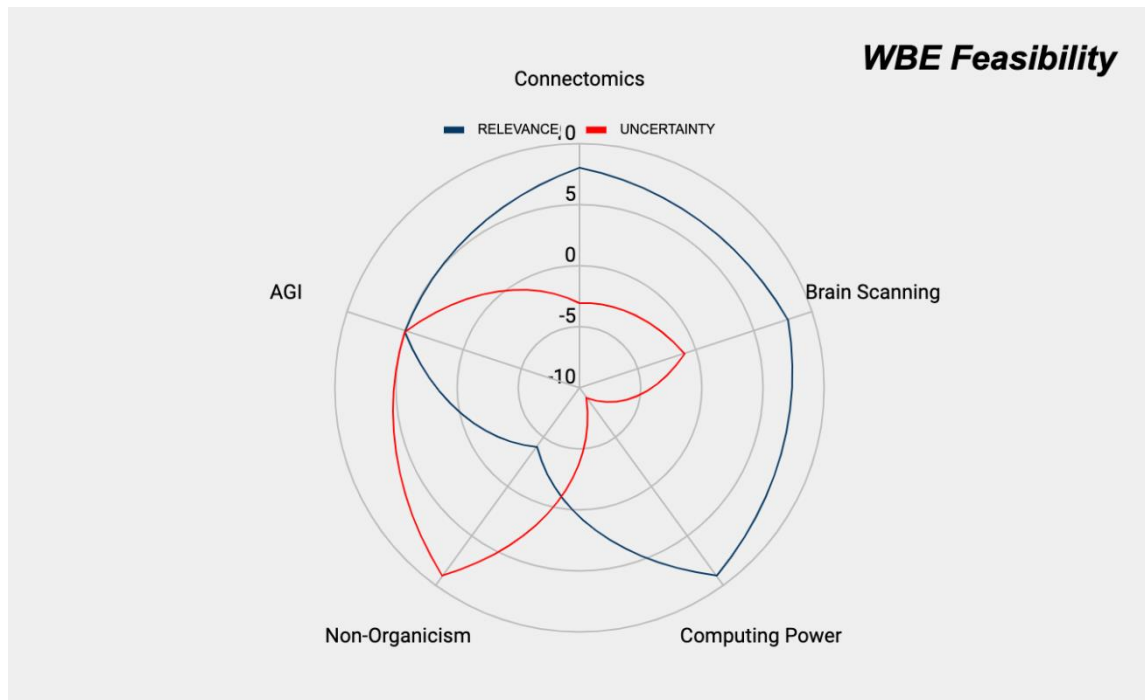


Figure 32 – Connectomics, Brain-Scanning and Computing Power quite reliable, whereas AGI and Non-Organicism are quite uncertain regarding the path towards a WBE feasibility. (Luig, 2020).

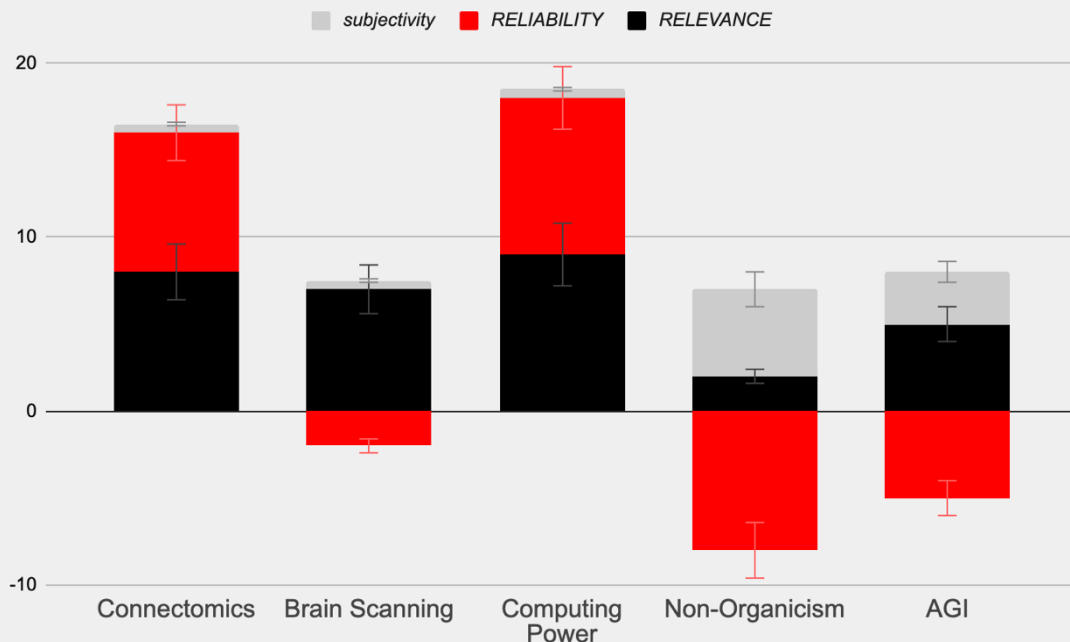


Figure 07 – Plot B (*WBE Feasibility*) focused on the subjectivity relative to each critical issue.

Table 20 – Plot B *WBE Feasibility* data.

CRITICAL ISSUE	RELEVANCE	RELIABILITY	Subjectivity level	(relative to each other)
Connectomics	8	8	relatively subjective	25%
Brain Scanning	7	-2	relatively subjective	25%
Computing Power	9	9	relatively subjective	25%
Non-Organicism	2	-8	absolutely subjective	75%
AGI	5	-5	subjective	50%

Regarding the levels of subjectivity in attributing weighed values:

Absolutely subjective _ meaning my personal view would be strongly opinionated or without an actual reference (example - an ideological position or taste in coffee);

Relatively subjective - meaning most people would think likewise (example - the earth is a round object, there is sufficient proof to believe so).

The best answer to the question, “Will computers ever be as smart as humans?” is probably “Yes, but only briefly.” —Vernor Vinge

