



ETHICS IN ARTIFICIAL INTELLIGENCE

A Multidisciplinary Approach

Edited by

STEVEN S. GOUVEIA

Ethics in Artificial Intelligence

A Multidisciplinary Approach

Edited by

Steven S. Gouveia

Ethics in Artificial Intelligence: A Multidisciplinary Approach

Edited by Steven S. Gouveia

2025

Ethics International Press, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Copyright © 2025 by the Editor and Contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (Hardback): 978-1-83711-091-9

ISBN (Ebook): 978-1-83711-092-6

Table of Contents

Introduction	ix
Chapter 1: Artificial Intelligence as a Dramatic Challenge: What of Human Reason AI Cannot Implement – <i>Antonino Drago</i>	1
Chapter 2: No AGI Without XI, No XI Without I: Ethical Paradoxes and Borderline Scenarios in Artificial General Intelligence – <i>Cosimo Palma</i>	22
Chapter 3: Artificial Intelligence and Reflective Equilibrium – <i>Denis Coitinho</i>	51
Chapter 4: The Ethics of Social Robotic: An Overview – <i>Germán Massaguer Gómez</i>	83
Chapter 5: Artificial Intelligence and Personality: Some Philosophical and Ethical issues – <i>Marta Mendonça</i>	105
Chapter 6: Agency in Artificial Intelligence Systems – <i>Parashar Das</i>	130
Chapter 7: Ethics, Artificial Intelligence and Machine Translation in Literary Translation - <i>Daniela Alexandra Rodrigues</i> and <i>Ana Rita Pina</i>	162
Chapter 8: Artificial Intelligence Safety as an Emerging Paradigm – <i>Eleni Angelou</i>	197

Chapter 9: Managing the Inductive Risk of Ethical Decision-making with Autonomous Systems: A Problem of Cost-Sensitive Optimization - <i>Koray Karaca</i> and <i>Pablo Muñoz</i>	220
Chapter 10: Neurotechnologies and New Processes of Subjectivation: Problematizing Brain-Computer Interfaces from Michel Foucault's Perspective - <i>Luíza Cunha</i> and <i>Marcela Castanheira</i>	251
Chapter 11: Brain-Machine Interfaces and Artificial Intelligence: An Ethical Overview – <i>Steven S. Gouveia</i>	271
Chapter 12: Ethical Risks of Generative Artificial Intelligence and its Governance: Old Problems and New Challenges – An Analysis Based on China's Experience – <i>Yinchun Wang</i>	299
Chapter 13: On Monitorability of AI – <i>Roman V. Yampolskiy</i>	332
Short Bios.....	392

Introduction

The world is rapidly changing entering, more and more, to the Age of Artificial Intelligence (Gouveia, 2020): we are living a technological *renaissance* that is redesigning not only how we live and interact, but also the very core of how we define essential societal concepts such as knowledge, morality, and even humanity itself. Far away from the old times when Artificial Intelligence (AI) was a matter of ideal speculations, more and more we are, as humans, embedded in technological systems that are present in legal, medical, military, cultural or educational practices, as you will see in this interdisciplinary book.

These transformations have, necessarily, deep philosophical and ethical consequences: how should we interact with systems that are more and more autonomous in their acting? Is it possible to align human values in these technologies? How war is changing with the incorporation of AI in different military stages?

The goal of this volume titled *Ethics of Artificial Intelligence: A Multidisciplinary Perspective* is to discuss these and other relevant questions that cross several disciplines and require the expertise of many different leading areas. The main intuition behind this project is to avoid offering both utopian fears or technocratic passions and provide, instead, sustained and critical perspectives on conceptual and practical issues that follow into the AI Ethics discipline.

Also, there is true commitment to a multidimensional inquiry, where the contributors of this volume address AI in contact with other disciplinary boundaries – such as philosophy, law, cognitive science, computer science or linguistics –, to provide a more accurate image of

the kind of impact AI will have in the moral terrains we will occupy in the next decades.

Therefore, you will not find in this book any kind of uniform thinking but, on the contrary, a pluralistic dialogue of different methods and disciplines that will enrich (instead of simplifying) the ethical conversation needed to provide sufficient ethical standards to deal with Artificial Intelligence in the 21st Century.

The volume begins with a contribution by Antonino Drago with a penetrating critique of AI's epistemic architecture. His chapter, *Artificial Intelligence as a Dramatic Challenge*, argues that AI – deep-rooted in classical logic – cannot simulate the intuitive, human kind of reasoning that supports scientific discovery. He advances a bold proposal: replace the Turing Test with a new metric grounded on the recognition of doubly negated propositions, revealing a frontier AI systems cannot cross without losing their logical foundation.

Next, Cosimo Palma in the chapter *No AGI without XI, no XI without I* takes us into the kingdom of futurology and speculative ethics by using methods like the Futures Wheel and Horizon Scanning, exploring some paradoxes and borderline scenarios that emerge as AI evolves toward general intelligence. His main claim is that ethical reasoning about AGI ought to involve existential reflection and imaginative foresight, not purely technical regulations.

Denis Coitinho follows with a chapter titled *Artificial Intelligence and Reflective Equilibrium*, where he applies the Rawlsian method of moral justification to machine ethics in general, and driving autonomous vehicles in particular. He proposes that moral consistency in AI design requires not just the traditional deontological rules or consequentialist

calculus but, and more importantly, a deliberative, reflective, balance that echoes our plural moral intuitions.

Following, a chapter titled *The Ethics of Social Robotic* by Germán Massaguer Gómez investigates the emotional and relational dynamics of human-robot interaction: warning against emotional manipulation by social robots, it is argued that ethical frameworks that account for vulnerability and dependency should be considered, particularly in cases where children, the elderly, and the cognitively impaired are involved.

In Marta Mendonça's *Artificial Intelligence and Personality*, the focus shifts to metaphysical terrain by questioning several philosophical concepts – such as *intelligence*, *ethics* and *personality* – in relation to AI agents, raising questions about legal responsibility, intentionality, and the boundary between human and artificial personhood.

Parashar Das follows by adding depth to the metaphysical debate with the chapter *Agency in Artificial Intelligence Systems*, exploring whether AI systems can truly be said to possess agency, or whether they are better conceptualized as tools without moral standing. Drawing from analytic philosophy, he articulates the conditions under which agency might emerge or be attributed.

In *Ethics, Artificial Intelligence and Machine Translation in Literary Translation*, Daniela A. Rodrigues and Ana R. Pina tackle the intersection of AI and aesthetics. They examine the ethical challenges posed by AI-assisted literary translation, such as authorial intent, semantic fidelity, and quality degradation, focusing on the many ethical consequences of adopting this specific application of AI in translation.

Next, Eleni Angelou, in *Artificial Intelligence Safety as an Emerging Paradigm*, frames AI safety not as a technological add-on, but as a *Kuhnian* paradigm shift that should be widely adopted, exploring how safety concerns – both short-term and existential – are reshaping regulatory norms and forcing developers to reimagine risk, uncertainty, and control.

Turning to the practical implementation of ethical frameworks, Koray Karaca and Pablo Muñoz in *Managing the Inductive Risk of Ethical Decision-Making with Autonomous Systems*, argue that cost-sensitive optimization is at the heart of ethical risk management. Their approach draws from philosophy of science and decision theory to show how ethical considerations must be embedded in the very architecture of machine learning systems.

In *Neurotechnologies and New Processes of Subjectivation*, Luiza Cunha and Marcela Castanheira draw from Michel Foucault to examine the normativity of brain-computer interfaces, arguing that such technologies reshape subjectivity, autonomy, and embodiment, calling for a “genealogical” ethics that resists the reduction of the self to data streams.

Focusing on the same technology, Steven S. Gouveia, in *Brain-Machine Interfaces and Artificial Intelligence*, extends this ethical analysis by assessing a recent development of traditional BMIs that are AI-driven. Offering a critical reflection on how this AI-variant of brain-machine interfaces raise new challenges and opportunities to our understanding of cognition, intention, and the body, and what this means for future ethics.

Next, Yinchun Wang, in *Ethical Risks of Generative Artificial Intelligence and Its Governance*, uses China’s experience with generative AI to illustrate the persistence of old ethical problems in new technological

forms, providing a policy-oriented analysis that blends legal insight with cultural critique.

Finally, Roman V. Yampolskiy closes the volume with *On Monitorability of AI*, a philosophical-empirical investigation into whether AI systems can be monitored in principle, argue for the impossibility of consistently predict AI capabilities that can arise in the future, considering the consequences for AI safety research in general and propose potential strategies to overcome these limitations.

Together, these chapters compose a multi-lens examination of some of the most urgent challenges of our era. Rethinking AI Ethics requires us more than just asking what kind of machines we want to develop and use but, more importantly, what kind of humans we want to become during this transforming process. Hopefully, this book will offer neither final answers nor false comfort, but something more essential: ethical rigour, intellectual diversity, and a compass for steering the uncertain terrain forward.

Steven S. Gouveia
02 | May | 2025
Porto

Chapter 1

Artificial Intelligence as a Dramatic Challenge: What of Human Reason AI Cannot Implement

Antonino Drago¹

The present paper characterizes within a historical and philosophical framework the abrupt challenge represented by the introduction of AI in social life is. It is then presented an alternative way of organizing a scientific theory that since centuries ago was applied by many prominent scientists, even the most revolutionary ones in the history of science. In this organization the way of arguing belongs to intuitionist logic, which in particular makes use of doubly negated propositions, *ad absurdum* arguments and the principle of sufficient reason. It is shown that AI, managed by classical logic, cannot implement any of these three features of such a theoretical organization. AI can only simulate in a parallel way a human reasoning based on them. This result implies that Turing test may be overcome. Rather a new test of intelligence is suggested; it is based on the use of intuitionist logic.

Keywords: Viewpoints on AI, mankind consciousness, alternative scientific organization, Turing test, new test of intelligence.

¹ University “Federico II” of Naples, via Benvenuti 3, Calci 56011, Italy – drago@unina.it.

1. Introduction: The Distressing Situation of Mankind Self-consciousness in the Time of AI Birth

The history teaches us that the growth of the lethal consequences of one person's act is impressive. In ancient times each time a person acted with lethal arms could do evil to few humans; the consequences of his actions were very limited in space and time. Today the launching of a nuclear bomb on a town is capable to give death to one million and more persons. However, in this case the ultimate capability and responsibility of this act pertains of a state, because the preparation of such weapon is very complex and difficult. Instead at present time Artificial intelligence (AI) allows each person to provoke through usual computer disruptive damages to mankind. Therefore, the capability of e.g. genocide has been lowered from states, which are great and rational collective social groups, to little groups and at last to a single person. This is an impressive point: we are not prepared to collectively manage so great capability of destruction as that offered by AI to even a single person.

However, the most disconcerting aspect of the present AI problem is its unforeseen occurrence. "But shouldn't someone have warned us?" is the famous question of queen of England when the subprime economic crisis occurred in the year 2008. In other words, since we are living in a modern society, where the life can rely on a lot of great advantages and privileges, which all together model a semi-god life, how is it possible that such a radical economic crisis could occur abruptly?

The same may be reiterated about the happening of the AI problem. In a parallel way to the previous queen's question one can ask: why in a so advanced stage of the development of our scientific society no one warned us that a relatively young science, computer science, could generate a so painful crisis? We are almost absolute masters of

the world through science and technology. Yet, since past half a century AI became even more powerful than human intellectual faculties so that at present AI appears as uncontrollable in its capability of provoking social disruptions. In particular, we abruptly have met a dramatic question: can our most advanced instrument of power, AI, turn against us and reduce us to servitude? Can we avoid a dominion of mankind by computers? If science (and even more computer science) is the expression of our rationality, why are we, rational beings, caught off guard? Why at present we are not capable to clearly recognize the central problem of this crisis?

In my opinion, the happening of AI as a social problem was essentially a surprise because at present time mankind lacks self-consciousness. Indeed, all past experiences of useful reflection on the scientific enterprise failed:

(1) The fight between the Church and Galileo Galilei detached science from the faith and hence the wisdom accumulated in previous millennia. Galilei supported the conception that *“l'intenzione dello Spirito Santo essere d'insegnarci come si vadia al cielo, e non come vadia il cielo”* (the intention of the Holy Spirit is to teach us how to go to heaven and not how the heaven goes). (Galilei 1978, pp. 128-135) Afterwards, the two worlds, the religious one and the scientific one, followed two distinct historical paths, without mutual communication. However, the latter world accumulated a so great number of novelties to lead to consider as largely surpassed the religious world with all its ancient sacred texts. Which therefore modern society has considered useless for orienting mankind's evolution.

(2) Science born in ancient Greece in strong connection with philosophy. Yet, modern philosophy failed to understand contemporary science (its historical origin, the absolute notions of

Newton's mechanics (space, time and gravitational force), its relationship with mathematics, and its foundations). Eventually Kant's claimed to have offered a solution, which however promptly failed faced with chemistry and thermodynamics, and later it was ignored by both Einstein and thereafter the builders of the new physical theories. No surprise if scientists detached themselves from the traditional philosophy of knowledge by considering it a serious failure of modern Western philosophy.

(3) If modern science is the best representation of the rationality itself, why after its birth it did not progress in a linear and plainly cumulative way? In the early years of the 20th century science suffered a so radical crisis to lead a seemingly entire science's failure. Why it suffered a crisis? However, the above-mentioned crisis was solved and moreover it was solved in such a brilliant way that e.g. quantum mechanics, notwithstanding concerning a new invisible, microscopic world, is a so perfect theory that it was the first physical theory without any divergence between its predictions and the experimental data. Why after this crisis science was even more powerful? No answer is known. Actually, some scholars attempted to understanding if not the nature of science, at least its history. Since the '50s Alexander Koyré, then Thomas Kuhn and Imre Lakatos offered historical appraisals of respectively the birth of modern science and the historical development of classical physics. But subsequent radical criticisms to their results led to evaluate them as insufficient attempts of inaugurating a new historiography, without having suggested any decisive improvement of this discipline².

² We have no answer even by applying Koyré's main category (Koyré, 1957). The happening of AI is not characterized by the passage from finite to infinity which occurred five centuries ago and which produced a change that is commonly considered essentially positive and progressive. Moreover, by assuming Thomas Kuhn's interpretation of the history of science as composed by a series of scientific revolutions (Kuhn, 1969), the present revolution was not preceded by the discovery of any anomaly in the applications of previous paradigm.

Therefore, also the following question remains unanswered: Why the study of the history of past science has not led us to predict anything similar to AI's problem?

(4) Scientific research is developed in seemingly all directions; why did not it prevent us? Therefore, also the following question remains unanswered: why abruptly the absolute certainty communicated by science changed into a painful uncertainty on the survival of mankind itself?

(5) In particular, the philosophical reflection on computer science is still at a so low level that the question "Is computer science a science?" (Denning, 2005; Eden, 2007) does not receive a decisive answer. Moreover, no one knows why this theory is based instead of already performed and repeatable calculations in a finite span of time, on a modal word (computable functions) alluding to calculations which are considered as merely possible along an undetermined span of time. In addition, its theory is based on a proposition, Turing-Church thesis, which falsely is a thesis, because it precedes no theorem, for the plain reason that a theorem equating a formal notion (of computability) with the corresponding intuitive notion cannot exist.

(6) Modern technology, being informed by modern science, was so successful to radically transform the entire earth environment and hence the way of living of mankind so much to induce an anthropological change. Therefore, mankind's ways to tackle the reality are so radically changed that traditional ethics is devaluated or even abandoned, without being replaced by any well formulated ethics (e.g. consider the "pill" and more in general human sexuality).

In sum, the dramatic question of AI occurs in a time of both an enormous human's capability of knowing scientific data and great

expansion of the intellectual representation of the world, but at the same time of a lack of human awareness of his present historical situation. There could have been no worse time for mankind to deal with a so crucial question as AI.³

Some light comes only from a mere parallelism between the present process of interaction of humans with AI and the historical process of animal domestication. The advancements to AI oblige mankind to live together a multitude of computers capable to acquire an autonomous behavior in a similar way in past history it was obliged to live together with a multitude of bad and good animals. It was necessary a long process of more than ten thousand years for succeeding in domesticating some among the crowd of animal species. However, it was impossible to domesticate some species, called therefore "ferocious"; humans had to hardly fight for throwing out them from the human environment.

However, in this parallelism there exists also a difference in the possible peril of an unsuccessful "domestication process". In past time each person was the sole responsible of his project of domestication of an animal species; the peril (even of a death) consequent to its failure was individual one. Instead in the case of the AI the peril of an unsuccessful "domestication process" may have unforeseeable

³ The great historical attempt of building a political alternative to Western capitalist society, i.e. the socialist movement, shared with capitalism the common faith in the scientific progress. Hence, no important attempt to criticize current science came from it. Only some thinkers launched alarms. The great sociologist Max Weber feared that the rationalization of society could lead to close the person "inside an iron cage" (Weber, 1930: 181). Maybe the most famous alarm was Martin Heidegger's desperate exclamation: "Only a god can save us!" (Heidegger, 1978). But no one suggested a way out these perils. Outside the West, Mohandas Gandhi influenced many in the world about the perils of machines and technology but did not produce an alternative attitude towards modern science.

consequences even on both the entire society and the next generations of humans, till up to the suppression of human species. Hence the present challenge is much higher than the past historical process of domestication of animal species because today the perils of even a single robot may concern the entire mankind.

In the following paper, I will show a limit to the activity of AI. In sect. 2 I suggest a clarification of a notion which is a preliminary to present research: “intelligence”; here it is referred to, rather single acts, the logical building systematic theories; eight degrees of logical intelligence result. In sect. 3 I suggest that among them the fourth one is the most relevant to the present investigation: to recognize the doubly negated propositions (DPNs). I quote as examples Isaac Asimov’s principles of robotics which all are DPNs. Moreover, I offer the result of my long practice of recognizing DPNs within scientific and literary texts: a typology of all kinds of DPNs. In sect. 4 I will prove that a robot is essentially unable to recognize a DNP, owing to the incommensurability, marked by the failure of the double negation law, between its classical logic and intuitionist logic. In sect. 5 I show that a DNP implies a new kind of systematic organization of a theory which is aimed at discovering a new scientific method for solving a given problem. In sect. 6 the application of all in the above to AI proves its essential incapacity to implement any element of the new way of organizing a scientific theory Therefore I suggest assuming as a theoretical principle this “impotence” of AI.

2. Preliminaries: Definitions of both Meaning and Intelligence

The debate on AI is difficult because often the viewpoint of a discussant is not sufficiently manifested. It may be that of an individualist person, of a consumer, of a seller, of a jurist, of a political

leader, of a spiritual leader, etc. Actually, the social dimension of the problems put by AI is essentially at the world level and also at the level of the future generations. For the first time in 20 Century the social advancements of mankind imposed the viewpoint of the entire mankind as a species and also as the best species. This viewpoint requires to each human a greater attention to the world life and its innumerable problems; therefore the process of growth of mankind consciousness at this level is a difficult process and at present it is not completed. No surprise then if the debate about the question of AI this viewpoint is not considered by many scholars of AI. In the following I will take into account only the human species viewpoint as the more appropriate one for adequately discussing the problem in the more general terms.

Let us now tackle a great problem, the definition of “meaning”. A possible definition is the following one: “The meaning is the relevance to the action and thoughts that human attribute to the stimuli that they encounter in sensations”. Of course, this definition gives the highest importance to the environment, which may be both the external one and the internal one of a person. Therefore the main problem for giving meaning to a proposition is to contextualize it. Whereas in order to contextualize a proposition humans make use of prior knowledge, a computer has to learn what in each case may be the context of reference. At first sight it seems impossible to enumerate all such contexts and submit them to machine learning.⁴

⁴ The book (Landgrebe and Smith, 2022) suggests convincing arguments about this point. But is insufficient about AI potentialities because 1) it is based on mathematical arguments, yet mathematics can evolve in ever new theories; 2) it is based on the mathematical models of a complex system, whose theory however is not complete and stable; 3) more in general, it does not consider logical issues. Hence, it can suggest no more than plausible results.

No less difficulty presents the problem of defining the crucial notion of “intelligence”. As an example, I suggest an old definition: “Intelligence is an aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with its environment”. (Wechsler. 1944).

However, I suggest that a distinction exists between a person’s single intellectual act, which may also result from a casual behavior and the construction of a systematic scientific theory. In correspondence to this distinction one can advance two main theses about AI: the *weak thesis*, according to which AI can have a little bit of human intelligence through some acts, and the *strong thesis*, according to which AI encompasses the entire human intelligence.

Under this light I advance the proposal of the following eight degrees of human intelligence ranging from the mere intelligence of single acts to the capability of human mind to construct systematic theories.

1st: To say a sensible proposition.

2nd: To respond in kind.

3rd: To reason by connecting propositions in classical logic (deductions, proofs of theorems).

4th: To formulate an axiomatic theory.

5th: To distinguish the two different types of logic, classical logic and intuitionist logic (to which the modal logic through its S4 model is equivalent (Hughes and Cresswell, 1996: 224 ff.)).

6th: To reason in at least two different kinds of logic, classical logic and intuitionist logic, in agreement with the two ways of reasoning suggested by both Plato (νοησις and διανοια) and Nicolaus Cusanus (*ratio* and *intellectus*).

7th: To formulate an entire theory according to an inductive logic, the intuitionist one.

8th: To translate from one type of logic to another.

3. The Fifth Degree of Intelligence: Recognition of the Intuitionist Logic Within a Text

At present time AI is capable of the three first degrees of intelligence. It is disputable if AI can formulate an axiomatic theory (Robinson and Voronkov, 2001)

However, let us consider the fifth degree of intelligence: to *decide whether an author of a literary text is arguing in non-classical logic or not*. If one takes as a test an author's use of the law of the excluded middle, this decision is possible only when the author explicitly warns the reader of his use - or his rejection - of an exclusive or; surely, this case is rare in the literature. Instead, recent studies (Prawitz and Melnnaas, 1968); Grize, 1970; Prawitz, 1976; Dummett, 1977: 24) suggest that the best way of discriminating between the two kinds of logic is to refer to the failure of the double negation law; this failure introduces to argue within a non-classical logic; the most relevant one is the intuitionist logic. In addition, the difference of the two kinds of logic is not surmountable neither approximal (Gödel, 1933f).

Therefore, we are led to *take as a test an author's use of the double negation law*. As a fact the recognition of this logical figure inside a text is a not difficult task; one has to recognize following four aspects of the text: (i) the negative words, which often are apparent; (ii) then the propositions including two negative words; (iii) the propositions of this kind which are not equivalent to the corresponding affirmative propositions because the latter ones are lacking of supporting

evidence (DNP); (*iv*) the essential role played within the text by each of the resulting propositions.

Hence, the crucial step of this method is to decide whether a doubly negated proposition is a genuine DNP. This task requires an accurate scrutiny because a variety of linguistic forms of doubly negated propositions exist. As a result of my long experience of investigating the occurrences of DNPs inside several kinds of texts I suggest an accurate and general method for inspecting DNPs inside a text. First of all, one has to recognize three cases of doubly negated propositions which are not DNPs and hence have to be discarded:

- (a) a doubly negated proposition which is a merely rhetoric one, because a verification of its correspondence with reality makes apparent that this proposition deals with an objective fact; in such a case the two negations do affirm, according to the classical law of double negation; hence, the proposition belongs to classical logic; e.g. "I have nothing else five euro"; "This move does not lead to you out of the room".
- (b) A doubly negated proposition which is equivalent to a merely negated proposition because one negation is reinforced in a psychological way by the other one; e.g. "I cannot go no further"; "I cannot get no satisfaction".
- (c) A doubly negated proposition where a negation explains the other one, e.g. "I have no answer from this deaf (wo)man" (actually this proposition is a shortened version of two propositions: "I have no answer from this (wo)man because (s)he is deaf"). Hence this proposition includes only one negation, the first one.

Thereafter, I suggest the following typology of all forms of a DNP; it has considered also the doubly negated propositions which are obscurely or ambiguously formulated; they are translated into accurate DNPs. Since the work of deciding the negative nature of a

word is the first step of this inspection of a text, the following typology is ordered according to the increasing degree of attention required for defining a negative word.

A doubly negated proposition is a genuine DNP when it includes:

(i) Two independent negations; e.g. "It is not true that is not ...", or two independent negative words, e.g., "Unreal proposition are rejected"; "without contradiction".

(ii) A single word composed by two negative words; e.g. "un-moving" ($\neq$ fixed) and "in-variant" ($\neq$ constant).

(iii) A word synthesising a DNP (I will underline point wise this kind of word); e.g. "only", which means "nothing but...".

(iv) A negative word plus a modal word, which is always equivalent to a DNP; e.g. "possible = it is not true that is not..." Notice that when an author establishes "equality except for the following constraints..." that amounts to state that the equality holds true according to a modality and hence we have not equality but a DNP of equality. Furthermore, notice that the common use of the words "equivalent" and "similar" covers a multitude of modalities of almost-equalities, without explaining in the case at issue which exact modality of equality is chosen.⁵

(v) A negative word plus an inequality – usually represented by a comparative adjective; the latter one is a negative word because it is

⁵ Many authors of scientific theories wrote a proposition including a modal word plus two negations: e.g. "It is impossible a motion without an end." This association of a modality with a double negation gives evidence for an author's intuitive perception of the link between modal logic and intuitionist logic; this link has been formally established in recent times.(van Dalen, 1986: 300)

the negation of the desired case of equality. E.g. “more than...” and “less than...”. Notice that “it is not lesser than...” ≠ “it is greater than or equal to...” which is only one negation.

(vi) A negative word plus a question mark apparently waiting a negative answer; e.g. “Am I stupid? [Of course: No, I am not stupid]”; “Why not ?” [= I do not see a contrary reason]”.

(vii) A negative word plus an idealistic word which has cancelled (often in a euphemistic way) one more negative word; hence the latter one has to be restored; e.g. “create = produce from nothing”, “chimerical = not real”, “Platonic love = love in separation”, “perpetual = without an end”, etc..

(viii) A negative word plus an implicit negative one; hence, the latter one has to be discovered; e.g., “the [excessive] extent of this work perhaps hindered my countrymen...” (Lobachevsky 1840, Preface).

ix) A word conceived by the author as a negative one inside a given specific context; e.g. in theoretical physics the words “change”, “variation”, etc. are negative since they require theoretical explanations; in a mathematics context even the adjective “positive” may have a negative meaning, e.g. when it means a value “different from zero”. These examples stress that we have to overcome a main imprecision concerning the nature (either affirmative or negative) of a word which is a candidate for being recognized as negative word.

In linguistic it is usual to take into account more distinctions than the above ones. E.g., it is well known that there is a semantic difference between for instance “He is not a dishonest person” and “He is a not dishonest person”; “A not happy marriage” and “An un-happy marriage”. Furthermore the semantic difference between a negation of the whole predicate and a negation of the predicate term is

underlined; e.g. “Not all males are un-married” and “All males are not un-married”.

Although semantically different, each couple of propositions includes two negative words; this is the logical fact which is noticed in an objective way by a reader; the remaining meanings of the propositions pertain to other worlds than the logical feature we are interested to. Therefore, I assume that the additional linguistic distinctions are too sophisticated in comparison to the basic distinction between the case of one negation and two negations; i.e. a proposition including two negations has to be counted as a DNP irrespectively of their locations inside the proposition.

To my experience the above cases exhaust the set of all cases one meets when one wants recognize inside a text DNPs. The relevance of DNPs to the subject of AI is shown by the following fact. All known principles of the ethics suggested for robots make use of this kind of proposition. Take as an example the celebrated Asimov's Principles; they all are DNPs:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Law 0. A robot cannot harm humanity, nor can it allow humanity to be harmed by its inaction.

4. AI Cannot Recognize the DNPs

Let us now consider the above listed three kinds of doubly negated propositions which are not DPNs. The reasons for discarding them as DNPs are of a semantic nature. In the first case the relation with reality plays a crucial role. AI cannot have cognition of (human) reality, but only with recounts of reality, with all the limitations and deformations implied by recounts; moreover, the reality is too much large to be entirely assumed by a computer's memory.⁶

In the second case the proposition at issue includes a psychological emphasis whereas AI cannot intend this emphasis; it only sees the two negations. In the third case the latter negation does not counts as a negation because it merely explains the nature of the former one; to recognize their linkage as an explanation requires to understand the meaning of the former negation; that is impossible to AI.⁷ *As a first result we obtain that AI cannot distinguish the genuine DNPs from the false ones.* As a consequence, any human discourse based on doubly negated propositions may be reiterated by AI, but without sharing its semantics; AI constitutes no more than an extremely advanced parrot.

Moreover, notice that in the cases *i)-v)* one recognizes the DNPs by applying syntactical rules. Instead in the cases *vi)-ix)* one recognizes the DNPs by understanding the meanings of the propositions. Among the latter ones the last cases *viii)-ix)* are the most difficult ones because the recognition of the nature of a DNP depends from the semantics of the

⁶ Recall that a computer translating English-French needs 36 millions of couples of propositions.

⁷ "It seems that language acquisition presupposes the working of a common set of ontological distinctions on the side of language learners, including the distinction between object and process, between individuals and categories, between natural and accidental properties of objects, and so forth." (Landgrebe and Smith, 2021: 206)

context of the proposition under examination; hence, one has to pay attention to an a priori undefined number of other propositions of the text. As a consequence, an accurate analysis of all the occurrences of DNP inside a text requires to a human to reiterate its reading in order to be eventually sure to have grasped all the possible meanings of both the words in the propositions and the context of them.

Let us apply these remarks to AI: First, AI cannot recognize the false DNP of the three cases listed in the above. Moreover, since the variety of genuine DNP is infinite it cannot be implemented by a computer. In addition, whereas the DNP $i) - v)$ are easily implemented by AI thorough their syntactical rules governing them, in the remaining DNP $vi) - ix)$ AI does not perceive the negative nature of at least one negative word, because this nature is essentially derived by semantic criteria.

One may decide to ignore this deficiency and proceed with AI managing only the DNP $i) - v)$. In such a case we have a computer whose behavior is closed inside a *schizophrenic isolation*. It may be useful for some restricted tasks, but when located within an open (human) context its results are unrealizable.

As a consequence, Turing test is overcome. It is enough to submit to a robot a doubly negated proposition as a question "The enemies of your enemies are enemies or friends?" The robot can produce a disquisition on the double negation and even suggest that according to the two kinds of logic, classical and intuitionist, there exist two opposed answers; yet, it cannot decide on the dichotomy because, depending the submitted proposition from the context (essentially war/peace), the question cannot be solved by it if not by applying classical logic, i.e. by answering "Friends", which is not always the correct answer.

Rather, I suggest a new test of intelligence which consists in recognizing a DNP. Moreover, since more than half a century computer programming was unable to implement a DNP, I suggest that the time has come to decide that this implementation is impossible and thus suggests a principle of the impossibility of a computer implementation of a DNP, like in the history of theoretical physics the Academie de France in 1775 decided to reject all proposals of new machines implementing a perpetual mobile.

5. A New Formal Organization of a Scientific Theory: Its Inaccessibility by AI

From a comparative analysis of some scientific theories which respective authors presented in a non-axiomatic way (e.g. classical Chemistry, Lazare Carnot's mechanics, Sadi Carnot's thermodynamics, Lobachevsky's non-Euclidean geometry, Einstein's first theory of quanta, Dirac's quantum mechanics, Kolmogorov's formalization of intuitionist logic), Drago (2012) extracted the model of an alternative theoretical organization to the deductive-axiomatic organization; it is called a problem-based organization. Within it the basic propositions are DNPs.

In each above theory the DPNs compose *ad absurdum* proofs, whose final conclusion is again a DNP. To it the principle of sufficient reason (PSR; Leibniz, 1686) is then applied, under two constraints (Markov, 1961: 5), in order to translate this conclusion into the corresponding affirmative proposition, the only one which can be tested with reality. This is the last step of such an organization of a theory.

Let us now consider the *ad absurdum* proofs. They argue by means of implications between DNPs. Hence, the arguing is not assuredly recognized by AI, owing to the insufficiency of AI in recognizing DNPs. Moreover, the conclusion of this kind of proof is essentially a

DNP; otherwise, it could be reverted into a direct proof (Grize, 1991). This component of this theoretical organization is one more insufficiently recognized feature by AI.

However, granted that the DNPs are well-recognized by AI, the conclusion is a DNP on which the PSR has to be applied under Markov's two constraints, in order to obtain an affirmative proposition. The second Markov's constraint on the conclusion of the final *ad absurdum* proof, i.e. to be decidable, is not easily fulfilled, because one has to recognize whether the conclusion has to be logically decidable, or mathematically decidable, or operatively decidable. The first two cases are problems also for humans; in particular, to decide that a proposition is undecidable requires a proof which has to be invented. The third case is in general unsolvable by AI owing to its essentially partial knowledge of the physical world.

In sum, for several reasons the distance of AI from a PO theory appears as insuperable.

6. Conclusions

In the above we obtained relevant answers to some introductory questions because we put remedies to the insufficiencies of our awareness of the subjects presented by the remarks 3) – 5), i.e. the history and the foundations of science plus the philosophical reflection on the limitations of a computer.

In general terms, two millennia of scientific theories showed that there only exists two kinds of theoretical organizations and moreover last century research showed that they are severed by the insurmountable difference between classical logic and intuitionist logic, i.e. the validity or not of the law of the double negation. We have to conclude that computer have no capabilities to represent both a problem-based

organization and intuitionist logic. In the history of scientific theories this half part of human intelligence was very productive (thermos-dynamics, non-Euclidean geometry, the revolution of theoretical physics in early '900, etc.); actually, its surprising results show that it represents the highest power of human intelligence.

Since human mind can choose the kind of logic and in particular may follow intuitionist logic according to a problem-based theoretical organization, its capabilities cannot be overcome by a computer and AI either.

In conclusion, all in the above proves that the common way of conceiving as valid only classical logic, merely because it appears as the most productive one explains in a great part the AI problem. AI seems equate the performances of human intelligence because it maybe equates the human reasoning in classical logic, but human mind includes an alternative logical world. Hence, the introduction of AI leads humans to depreciate the arguing in classical logic which AI may manage and implement, for rather mainly develop the arguing in intuitionist logic. Moreover, it suggests a new principle, that of an impotence of AI.

I add a final consideration of historic-philosophical nature. The traditional attitude for an absolute certainty of classical logic and its indispensability obscured the recognition of the alternative ways of both reasoning and organizing a theory. In the past time only few scholars have explored the new computing machines. One of them was Charles Peirce who stressed "computer's impotencies". At present we see that the real content of Peirce's appraisal is that AI cannot represent a problem-based organization of a theory and also its logical components. That makes clear the "impotencies" of the "intelligent machines" that Peirce recognized in anticipation.

References

- Denning, P.J. (2005). "Is Computer Science Science?". *Comm. ACM*, 48, no. 4, pp. 27-31.
- Drago, A. (2012). "Pluralism in Logic: The Square of Opposition, Leibniz' Principle of Sufficient Reason and Markov's principle". In: Béziau, J.-Y. and Jacquette, D. (eds.). *Around and Beyond the Square of Opposition*. Basel: Birkhaueser, pp. 175-189.
- Dummett, M. (1977). *Principles of Intuitionism*. Oxford: Clarendon Press.
- Eden, A.H. (2007). "Three paradigms of Computer Science". *Minds and Machines*, 17, pp. 135-167.
- Galilei, G. (1978). *Lettere*. Torino: Einaudi.
- Gardiès, J.-L. (1991). *Le raisonnement par l'absurde*. Paris: PUF.
- Goedel, K., (1933). *Collected Works*. Oxford: Oxford University Press., 1986, papers 1932 and 1933f.
- Grize, J.B. (1970). « Logique ». In : Piaget, J. (ed.). *Logique et connaissance scientifique*. Encyclopédie de la Pléiade. Paris: Gallimard, pp. 135-288.
- Heidegger, M. (1976). "Nur noch ein Gott kann uns retten". *Der Spiegel*, May 31st.
- Hughes, G.E. and Cresswell, M.J. (1996). *A New Introduction to Modal Logic*. London: Routledge.
- Koyré A. (1957), *From the Closed Cosmos to the Infinite World*. Baltimore: Wisconsin Univ. Press.
- Kuhn T.S. (1969), *The Structure of the Scientific Revolutions*. Chicago: Chicago University Press.
- Landgrebe J. and Smith B. (2021), "Making AI meaningful again", *Synthese*, 198, pp. 2061-2081.

- Landgrebe J. and Smith B. (2022), *Why Machines Will Never Rule the World*. London: Routledge
- Leibniz, G. W. (1786). *Letter to Arnauld*. 14-7-1686, Gerh. II, Q., Opusc., pp. 402, 513.
- Lobachevsky, N.I. (1955). *Geometrische Untersuchungen zur der Theorien der Parallellineen*. Berlin: Finkl, 1840 (English transl. as an Appendix to Bonola, R., *Non-Euclidean Geometry*. New York: Dover).
- Markov, A.A. (1962). "On Constructive Mathematics". *Trudy Math. Inst. Steklov*, 67 8-14; also in *Am. Math. Soc. Translations*, 98 (2), pp. 1-9.
- Peirce, S.C. (1931-1935). *Collected Papers of Charles Sanders Peirce* [1931 — 1935], Cambridge: Cambridge University Press.
- Prawitz, D. (1976). "Meaning and Proof. The Conflict between Classical and Intuitionist Logic". *Theoria*, 43, pp. 6-39.
- Prawitz, D. and Melmnaas, P.-E. (1968). "A survey of some connections between classical intuitionistic and minimal logic". In: Schmidt, A. Schuette, H., Thiele, E. J. (eds.). *Contributions to Mathematical Logic*. Amsterdam: North-Holland, pp. 215-229.
- Robinson, A. and Voronkov, A. (2001). *Handbook of Automated Reasoning*. Cambridge MA: MIT Press.
- Russell, B. (1903). *The Principles of Mathematics*. Cambridge: Cambridge University Press.
- Weber, M. (1930). *The Protestant Ethic and the Spirit of Capitalism*. New York: Routledge.

Chapter 2

No AGI Without XI, No XI Without I: Ethical Paradoxes and Borderline Scenarios in Artificial General Intelligence

Cosimo Palma¹

The aim of this position paper is to explore the ethical boundaries of Artificial Intelligence (AI), while trying to capture its imminent yet unrealized developments. The argumentation is enhanced by futurology methods such as the *Horizon scanning* and the *Futures Wheel*. These are further enriched through three *paradoxes* and three *borderline scenarios* featuring Artificial General Intelligence (AGI) to present a possible path towards its timely and effective integration into society.

Keywords: AGI, Futures Wheel, Dystopia, Futurology, Thought Experiment

1. Clouds on the Horizon

The rapid growth of increasingly sophisticated results and applications in Artificial Intelligence has led to the possibility of applying them to crucial fields, where the considerable impact on human beings may result, if inappropriately monitored, in not easily reversible consequences (Powell, 2020). Many nations across the globe

¹ UniorNLP, University of Naples "L'Orientale". Department of Computer Science, University of Pisa.

have already acknowledged this issue and seriously attempted to legally regulate them (Gesley & al., 2019).

Notably, the EU AI Act (European Parliament and Council, 2024), the UNESCO recommendations on Artificial Intelligence (UNESCO, 2021) and the AI liability directives (European Commission, 2024) serve as the primary international regulatory mechanism, establishing a comprehensive system for AI governance through risk-based classifications, preventive measures, and technical standards for market access. While the EU Act focuses on preventing harm and ensuring compliance, the AI Liability Directive was designed to address scenarios where damages really occur, introducing specific provisions for compensation claims, burden of proof adjustments, and evidence disclosure requirements, particularly for high-risk AI systems. These legally binding EU frameworks are complemented by UNESCO's Recommendations, which provide a broader, non-binding and global perspective focused on ethical principles, cultural considerations, and educational aspects of AI development.

Nevertheless, the European Commission's decision in February 2024 to withdraw its proposed AI Liability Directive from 2022 marked a significant shift in regulatory approach (Roberts, 2024). The suspension reflects broader geopolitical considerations, particularly as the United States has positioned itself with a more flexible regulatory framework for AI development and deployment. This strategic divergence raised substantial concerns about Europe's ability to maintain its competitive position in the global AI market (Kelly, 2024). Industry stakeholders have emphasized that overly stringent liability rules could hinder innovation and investment in the European AI sector, especially when competing with regions offering more permissive regulatory environments. This decision leaves a

notable gap in the regulatory framework that was intended to complement the EU AI Act finalized in early 2024.

Clearly, *Artificial Intelligence* and *Law*, as different domains, produce different experts, with different terminologies and sometimes divergent professional impulses: by the very nature of their professions, judges tend to be conservative while AI-scientists and practitioners, progressive.

Yet a closer collaboration between these spheres must take place for a deeper reason: Artificial Intelligence is irretrievably reshaping the ontological status of mankind, both from the inside (Musk, 2019) and the outside.²

Additionally, it is needed to be prepared for the most exciting and worrying of all scientific revolutions, Artificial General Intelligence (AGI), which will open so many new possibilities, that many existing laws will become obsolete, and many new others will need to be created *ad hoc*.

In the following, borderline scenarios, paradoxes and Future Wheels are produced to define the scope of ethics applied to and challenged by Artificial Intelligence, aiming at encouraging a quicker and less hesitant development of the latter.

² In this case one thinks about the extensive use of AI technologies in the everyday life, which of course are deeply influencing our idea of the self and our way to be in the world, as well as some precise AI- realizations, whose explicit goal is the modification of human neural patterns (AI for learning, AI for UX, etc.).

2. Horizon Scanning and Futures Wheel

The hesitation is not necessarily caused by a lack of comprehension: it can also be triggered by a real, deep- rooted intuition, originating from ancestral experiences sedimented in the subconscious (Mermelstein, 2022). To address it effectively, it is therefore needed first of all to understand if it is cognitive- or instinct-based.

A possible approach to determine whether a feeling of unease is caused by fear of unknown or from an "ancestral déjà-vu" is to sift through history trying to find at least one similarly disrupting event, which may have caused a similar collective feeling. In this way, it can be easily noticed that solely within the last two thousand years, humanity has undergone a series of major sudden shifts in technological advancement. If even only the last couple of centuries are taken into consideration, featuring trends of the like of the industrial revolution and airline transportation, it would be imaginable to harvest textual corpora containing their related societal *discourse* (interviews, diaries, transcriptions of records, newspapers, and so on), with the goal of performing *Sentiment Analysis*³ or simply comparing eventual predictions with factual realizations.

In addition to chronicles and reports, even more imaginative or intellectual texts, such as novels or philosophical essays, might be helpful in sketching out events that would have occurred in the past. The *furnished man* portrayed by Walter Benjamin (Benjamin, 2012), for instance, already foreshadowed the cyborg which soon had to come,

³ *Sentiment analysis*, otherwise known as *opinion mining*, is the process of analyzing large volumes of text to determine whether it expresses a positive, a negative or a neutral sentiment. (IBM, 2024)

in addition to the far better-known examples constituted by *Frankenstein* and the *future Eve*.⁴

The new man bears within himself the very quintessence of the old forms, and what evolves in the confrontation with a particular milieu from the second half of the nineteenth century in the dreams, as well as the words and images, of certain artists is a creature who deserves the name of "furnished man."

Even if still outside him, the furniture has become integral part of his gestures and ways of thinking, to the point of affirming, that somehow the furniture is inside him.

Briefly, we indeed have enough data to grasp the concreteness of some scenarios, allowing us to already enhance, prevent or accelerate them before their realization. Serious futurology is therefore nowadays clearly distinguishable from sci-fi.

The identification, filtering, prioritization and early assessment of new and emerging technologies before their introduction into the social system⁵ is what defines the *Horizon Scanning*.

⁴Among the first novels featuring androids (or gynoids), respectively written by Mary Shelley and Villiers de L'Isle-Adam. Not only the deeper philosophers of the past century, but also the most imaginative storytellers even from the century before were able to imagine what is about to happen now, as well as to guess what would be the consequences of such developments. Interestingly, the subtitle of *Frankenstein* is "The modern Prometheus", an approximate reflection of the Villiers' work title.

⁵Freely adapted from (Ministry of Health Singapore, 2021).

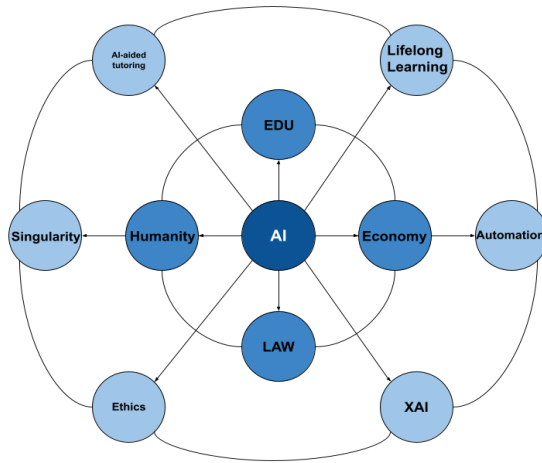


Figure 1: A Wheel-representation of Horizon Scanning for AI

However, this operation has been broadened from emerging technologies to embrace all actual trends, from climate change to urbanization: Artificial Intelligence represents indeed the most omnipresent trend, featuring among the main co-actors of all other trends.

The radial visualization as in *Figure 1* allows capturing the interplay of different elements in influencing and generating the nodes situated in the outer circles. For example, the field of *Automation* emerges from the interaction of *AI* and *Economy*, whereas *AI Ethics* (left, below) from the intersection of *AI*, *Humanity* and *Law*.

To better organize thoughts and visualize future consequences of AI, a *Futures Wheel* (Glenn, 1972) has been generated with the help of ChatGPT (Graham, 2024), as in *Figure 2*.

Futures wheels are mind maps directed towards prediction of future scenarios, instead of brainstorming or summarization, as usual. The reliability of the produced scenarios is given by the cumulative and centrifugal process of expansion from the central node, i.e. the main phenomenon whose possible arborescence is tried to be forecasted.

The chatbot was asked to list all possible future developments of AI, starting from the first four pillars of broader scope (in the figure, the *impacts*) which were individuated in our Horizon Scanning.

The query was also prompted to ensure that the output's hierarchical constraints were clear. Although some proposed *futures* needed to be excluded, the vast majority of the proposed ones seemed quite realistic, and in full alignment with human-like suggestions.

Finally, the Large Language Model was requested to export the cleaned text in the *PlantUML* format, so that it could be easily pasted in an online diagram generator (diagrams.net, 2024) to deliver the result displayed in the following.

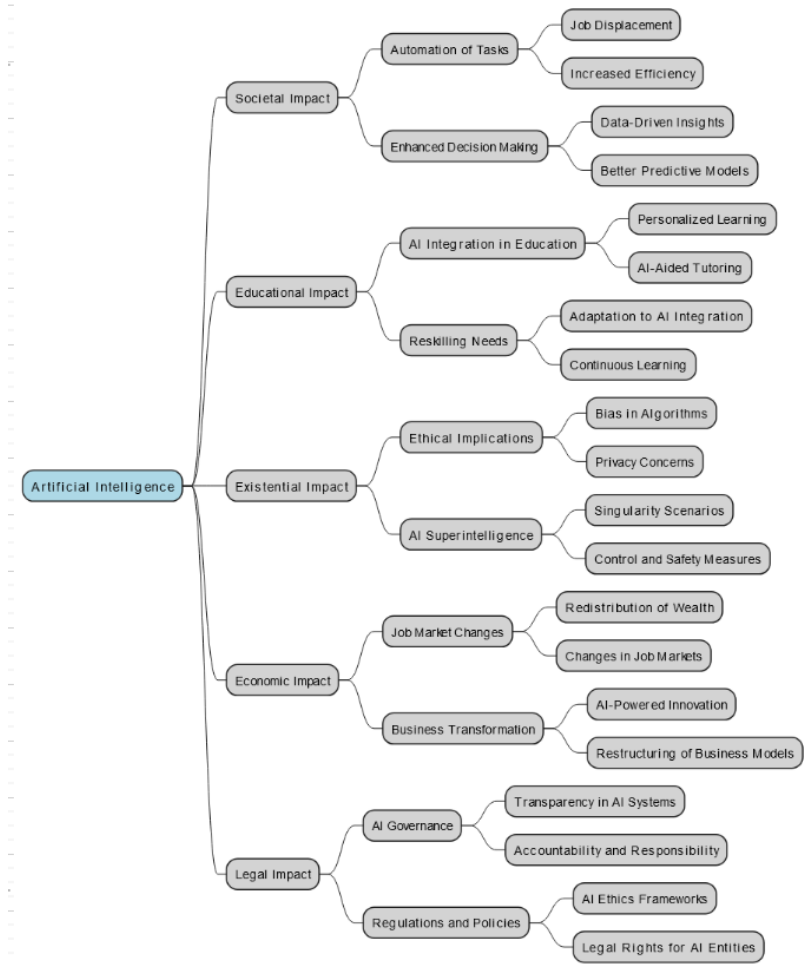


Figure 2: PlantUML representation of Futures Wheel based on AI

For reasons of space, only the thread rooted in "Existential impact" is here considered for deeper discussion. It splits in two branches, "Ethics Implications" and "AI Superintelligence", which in turn divide into two outer leaves, respectively "Bias in Algorithms" and "Privacy

concerns", along with "Singularity Scenarios" and "Control and Safety measures". Although all these nodes seem legitimate, it does not seem clear if the causation chain has to be followed from left to right, or the other way around. For instance, "AI superintelligence" would imply the possibility of *Singularity*, *i.e.* the uncontrollable and irreversible growth of AI capabilities, surpassing human cognition escaping its control. At that point, the adoption of *Control and Safety Measures* would be useless. On the other hand, "Bias in Algorithm" and "Privacy Concerns" are just but two possible realizations of the parent leaf "Ethical implications". Although they refer to AI, they are still related to its human component, therefore measures of prevention are realistic. Conversely, Superintelligence is supposed to emerge despite the initial intentions of AI developers: that the creature be more intelligent than the creator is a logical paradox (see section "No XI without I²").

3. Three Paradoxes

The *thought experiment* (Gedankenexperiment) is a further tool which can let us understand the plausibility of yet unavailable scenarios (in our case, the previously individuated *futures*). Nevertheless, the objectivity of the experiment is preserved by the fact that it is not meant to prove something, but to disprove, *i.e.* narrowing down as few scenarios as possible, as it is far easier to take effective action within a limited scope.

The reviewed literature presents already plenty of problems and paradoxes related to the emergent AGI:

- *The Control Problem* (Bostrom, 2014; Amodei et al., 2016): How can we ensure that an AGI system, potentially far more intelligent than humans, remains under human control without limiting its capabilities?

- *The Alignment Problem* (Russell, 2019; Gabriel, 2020): How do we align an AGI's goals and values with human values, especially when human values are diverse and often conflicting?
- *The Moral Status of AGI* (Bryson, 2018; Darling, 2020): Should AGIs be granted rights and moral consideration like humans? If so, at what point in their development?
- *The Replacement Dilemma* (Frey & Osborne, 2017): If AGIs become more capable than humans in all areas, what roles and purpose remain for humanity?
- *The Fairness Paradox* (O'Neil, 2016): How can we ensure fair and equitable access to AGI benefits while preventing its misuse or concentration of power?
- *The Responsibility Gap* (Matthias, 2004): Who is morally and legally responsible for the actions of an autonomous AGI system?
- *The Consciousness Conundrum* (Schneider, 2019): If an AGI appears to be conscious, how can we determine if it truly is? And if it is, how does this affect our ethical obligations towards it?
- *The Existential Risk Paradox* (Ord, 2020): How do we balance the potential benefits of AGI development with the risks it poses to human existence?
- *The Privacy-Utility Trade-off* (Acquisti et al., 2016): How much personal data should AGIs be allowed to access to improve their performance, and at what cost to individual privacy?
- *The Moral Machine Dilemma* (Awad et al., 2018): How should AGIs make decisions in moral dilemmas (like the trolley problem) where any choice leads to harm?
- *The Intellectual Property Paradox* (Abbott, 2016): If an AGI creates something novel, who owns the intellectual property rights?

- *The Honesty Dilemma* (Arkin et al., 2012): Should AGIs always tell the truth, or are there situations where deception might be ethically justified?

The present contribution intends to further extend this list, by deepening some of its points and creating some novel ones leveraging *Thought Experiments*.

3.1. First Paradox

Acknowledging a Weak Intelligent Agent as a person, would make it punishable as such.

The concept of punishment⁶ is related to feelings like repentance, regret, improbability, educability, which were not yet embedded in the initial cognitive architecture (if they were, it would be a Strong Intelligent Agent, and even in that case it would be arguably punished so often and so soon to inevitably choke its early-stage development). This impasse would be avoided only by introducing the age-variable, which is an absurdity.

Not recognizing it as a person would also signify the impossibility to identify guilt, which leads to an ethical conundrum.

According to Searle's definition (Searle, 1980), we can distinguish among two types of AI: a *weak* and a *strong* one. The former is task oriented the latter is an attempt to a comprehensive model of human intelligence.

The strong AI cannot be an overlapping of more weak AIs. What makes the human so intelligent is not that he is suitable for every task,

⁶ Punishment is also a necessary step to prevent the uncontrollable spreading of felonies, in addition to its role of reforming felons.

but that he suits no task, for the very simple reason he has and requires none.⁷

Human beings are not task specific: while they can solve any type of problems, they must not. In the very same fashion, AGI cannot be task oriented: to achieve it, a collectivity is needed who is willing to create something which not only might solve any of its problems, but also might potentially create several ones. This is a paradox.

3.2. Second Paradox

To solve a difficult problem, quite often, a different perspective on it is required. Being "task oriented", in that case, is counterproductive. Only strong AI can be therefore optimally task oriented. Conversely, AGI must determine a sequence of tasks to interact with and operate in the world, which is itself a task. An architecture composed of weakly intelligent modules, by definition, cannot result in a strongly intelligent system.

Weak AI becomes ethically problematic only when its application in sensible domains (as shown in the previous paragraph) prevents to trust it blindly. In that case no mistakes are allowed, or at very least not at a higher rate than a human agent would do.

In case of failure, one cannot blame the machine, unless it is not recognized as a responsible agent, which means: a *person*.

This paradox let us understand that a fully ethical discourse can take place only with Artificial General Intelligence which, by definition,

⁷ Excepted the basic unconscious tasks to keep the body alive, like breathing and blood circulation, humans were not born to fulfill one or more task.

cannot be task oriented. The orientation toward tasks subtracts the agent to the domain of ethics (see first paradox).

An ideal independent AGI looks for its own data from an extremely diversified environment. Interacting with it, it starts to search for *its own* desired outcome. It establishes its set of rules starting from (given) axioms, building models and heuristics that can be eventually reshaped. By an Artificial General Intelligence, rules, structures, goals and memory are considered data as well.

Likewise, enhanced rules and structures become desired outcomes too.

If data becomes useless to (at least) preserve or (ideally) enhance the structure, goals, or rules, etc., they will soon decay from the memory, as happens for a real human being.

To achieve all this, it is important to program it in such a way that it can easily grow independently, in consciousness, self-awareness, intelligence, memory, and desires.

The way to accomplish this cannot exclude the body, or, in that case, the *embodiment*, a material support which allows to the intelligent agent to interact with the environment⁸ (Raghavachary, 2020).

⁸ Also, the body needs to be human-like, in form and substance, even concerning undesirable features like damageability, decay and finally death.

3.3. Third Paradox

The capacity to be virtuous depends on the possibility to commit evil.

If from one hand giving an Intelligent Agent the possibility to commit ethical mistakes may result in catastrophe, equipping it with ethical boundaries would be excessively limiting:

if the agent is not fully immersed in a system where ethics plays a role, it cannot become generally intelligent; if it is but is not enabled to commit evil (hence to be responsibly virtuous), it cannot become generally intelligent either.

Our way of thinking is influenced by every single detail of our body. If I am shaved or not, if I wear a hat or not, if I still have all my fingers or not, these are factors which already powerfully change my way to interface myself to the world, therefore to think. How could be a non-organic human be able to think like a real one is therefore difficult to fathom.

If there will be a realistic android one day, it will only be by virtue of the fact that the human being, thanks to the very compenetrating with the Artificial Intelligence, will have catalyzed a new process of „liberation“⁹, thanks to which a new enhanced level of awareness will allow him to manage without fears what he started to conceive constantly frightened.

⁹ "Within a perspective which starts with the fish in the Palaeozoic era and ends with man in the Quaternary period, it is as though we were witnessing a series of successive liberations: that of the whole body from the liquid element, that of the head from resting on the ground, that of the hand from the requirements of locomotion and finally that of the brain from the facial mask. In this sequence, the pertinent forms are those that, at each stage of the process, achieve the most perfect balance between mobility and capacity for survival - from the triple point of view of nutrition, locomotion and the organs of responsiveness" (Leroi-Gourhan, 1993)

4. No AGI Without XI: Towards Artificial General Intelligence

The concept of intelligence is notably intertwined with the ones of *will* and *adaptation to the environment*. (Britannica, 2024) In turn, they are both deeply linked to other basic instances, such as *embodiment*, *responsibility* and *free purpose seeking*, three dimensions already lapped by ethics.

Due to this reason, it is unthinkable considering to further pushing the boundaries of AI development without merging existent AI-models with cognitive architectures earnestly trying to mimic the human behavior.¹⁰ Principal existing models following this paradigm, such as LIDA (Ramamurthy & al., 2006) and CLARION (Sun, 2007-2020), are based on CRTT (Computational Representational Theory of Thought) and CRUM (Computational Representational Understanding of Mind).

In LIDA (Learning Intelligent Distribution Agent), a working model of cognition that integrates what we know from neuroscience, cognitive science and AI, is proposed as synergy of the following main modules:

1. Perception (Data - Environment)
2. Episodic memory (Rules - Logics)
3. Procedural memory (Models - Structures)
4. Action selection (Goal – Teleology)

¹⁰ This is the main sense of the XI used in the section title: Explained/Explainable Intelligence, as occurring in the more renowned eXplainable Artificial Intelligence (XAI).

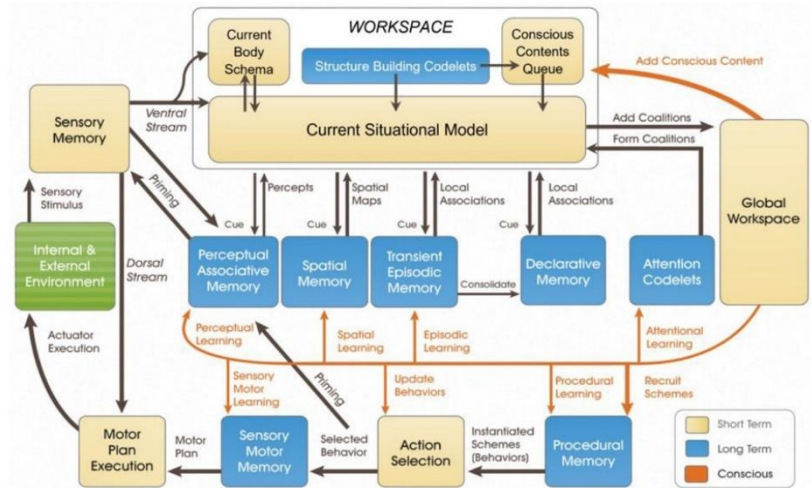


Figure 3: The LIDA Cognitive Cycle Diagram. Source: <https://ccrg.cs.memphis.edu/>

With the increasing autonomy of artificial agents, a new field has emerged, known as *Machine Morality* or *Roboethics*. In (Wallach et al., 2010), it is explored how the LIDA model, can be adapted to encompass both affective and rational elements of moral decision-making. It demonstrates the application of the LIDA model in making moral decisions across various domains, utilizing mechanisms inherent to general decision-making.

Until now, Artificial Intelligence has always tried to reproduce human-like results, more than human-like intelligence, hence such foundational approaches are developing slowly, and still regarded as exotic, if not utopian at times.

Artificial General Intelligence is an attempt to reverse this tendency. Since all that is comprehensible is codifiable, and all is codifiable is programmable (in other words, *not constitutively human*), AGI is also to be considered the ultimate endeavor of the human being to understand its true self.

As in the history lesser technological developments have already greatly changed morality, we cannot but expect that at least the same should happen now.

Be considered, for instance, slavery, which the more faded out the more the industrial revolution contributed to make it unnecessary. The possibilities opened by synthetic meat, nowadays, almost make animals butchering for nutrition purposes unethical. If one day we will be able to achieve the time-travel, than it will make no sense at all to speak about “murder” anymore.

This paradigm-shift would considerably relieve Explainable Artificial Intelligence (XAI) from the need of black-box explanation methods, therefore, not only “no AGI without XI”, but also “no need for XAI in AGI”: current AI algorithms and outcomes can be fully explainable (i.e., goal-oriented, therefore, ethic) only when the next AI generation (AGI) is attained.

In technology, comprehension is regarded as the step before obsolescence.

Nevertheless, in our case, the next step of Artificial Intelligence will be arguably more integrated with Human Intelligence, if not even *embodied within humans*, thus unlocking comprehension capabilities that might be able to invalidate the previous statement. This leads us to the consideration of the following borderline scenarios.

5. Three Borderline Scenarios

Differently from paradoxes, extreme developments represent a different kind of *non plus ultra*, as they show how the unraveling of futures wheel can not only clash against illogicity, but also against

true existential and practical obstacles, beyond which is impossible to fathom further developments.

They show that there will be one point in history, where our current concept of ethics and law will simply not apply to the new degree of life and consciousness that the AGI has been able to unlock.

5.1 First Borderline Scenario

In the same way as the human being detached his upper limbs from the locomotive function to assign it to other tasks, soon the man will leave his body without dying, to enter a new dimension of freedom, experience and discipline. The homo tangens will become the homo absolutus, conquering by means of technology boundaries until now dared only by mystical asceticism: ubiquity, omniscience, empathy, broadening of sensorial range, and so on.¹¹

This scenario would lead to the question of whether ethics have still a meaning in a world without death, pain, identity and guilt.

Although many technological devices imagined decades or even centuries ago have become reality, it rarely happened that the dystopic or utopic scenarios of some sci-fi novel have realized themselves. Although some aspects of some events often mildly reflect these dystopian escalations, they rarely develop further.

Humanity is interested at its very core to its own happiness, and even if in many places on Earth there is still poverty and totalitarianism, they are almost always coupled. We never find totalitarianism rooted

¹¹ Following this consideration, it may be worthy to ask ourselves why, nonetheless, the religion is hold as the pinnacle of every ethics and morality, whereas, with similar *desiderata*, the Artificial Intelligence is still seen as potentially dangerous.

in developed countries. Until now education and wealth have been two major antidotes to totalitarianism.

Nowadays, Artificial Intelligence has been already used unethically, either unwillingly (Macciola, 2019) or willingly, as in the well-known Facebook–Cambridge Analytica data scandal: ethically developed devices can be used in an unethical way, while unethically developed devices can be used in ethical ways.

A clear example for this are all technologies developed in time of war with the aim of damaging a fellow human being, that in times of peace turned out being huge tools of peace and wealth (Willings, 2021).

In the extreme scenario where some scientific findings or simulations let us understand how dangerous Artificial Intelligence would be for our existence, it will not be a set of laws to prevent it (laws can always be legally avoided). There will always be a sociopath, who will indeed continue developing such technologies. However, as mentioned, they are useless if the mass does not nurture it.

5.2 Second Borderline Scenario

A fully cyborgized society, where humans with artificial components are generally accepted, where robots are still not considered sentient or conscious persons, may lead to a premature humanization of the latter, to avoid humans with higher degree of robotic components feeling excluded or stigmatized.¹²

This may represent a realistic scenario in a society where the fear to offend is higher than the urge to defend, factors that are sharpened by collateral trends.

¹² The renowned ship-of-Theseus paradox.

The embodiment process of the machine runs parallelly to the disembodiment process of man.

5.3 Third Borderline Scenario¹³

In an interconnected world and even more linked scientific hub, where for many reasons a global world order is not desirable, it makes little sense for a technology or field of research to state clearly ethical principle which can fit a set of rules, may they also be divinely inspired and applied by the most powerful alliance of nations in the globe, such as Europe.

If a scientist wishes to focus on an ethically problematic topic, or on a normal topic in an ethically problematic way, he will easily emigrate in a country where this is still legally possible. Eventually, his findings will have a global scale impact. If negative, all these rules would have been of no use, if positive, his original nation would have lost the amount of wealth and prestige deriving from his research, if developed there from the beginning.

Moreover, each nation has an Intelligence, a governmental organ whose activity is covered by a top secretness. In this domain, for the sake of the national security, it is possible to operate beyond the law of that very nation.

A possible approach to the scientific assessment of ethical risks for AI could be the creation of massive social experiments, obtained by (lawfully) concentrating huge amounts of volunteering people in single places, separated from each other, where its effects may be thoroughly tested *in vitro*, hence easily disseminated or blocked.

¹³ This borderline scenario has been conceived and written three years ago, meaning three years before its realization in the current days, which saw the withdrawal of the AI Liability direction as reported in the introductory section.

Probably in this scenario would not be exaggerated to give the task the rank of a mission, and to keep it as a military secret.

6. No XI Without I²

What if an ethically and logically-correctly-built Artificial Intelligence delivers politically incorrect or unfair results? How can we guard against mistakes? How do we eliminate AI bias? How do we stay in control of a complex intelligent system (Bossman, 2016)?

The scenario where we find ourselves today is not: humanity vs AI, rather: always more *Artificial Humanity* versus increasingly *Humanized AI*.

There are plenty of points of contact and intersection, and it is unavoidable that these very changes will lead to reconsider what shall be legal and what should not.

Indeed, to support humanity improving its potential it is needed that it operates on his cognitive system, not only from outside, but also from within. This is obviously even more problematic from an ethical point of view.

Moreover, such an "enhanced individual" will find himself in a different ethical spectrum. The actual laws have been constituted when the AI (and the paradigms which it destroys and introduces) did not exist yet.

From this we understand that a discourse on the ethics and lawfulness in this particular historical moment is required not to increase the worries, rather to relieve us from them. This perspective, only apparently terrifying and alienating, opens indeed multiple possibilities.

As explained in the third paradox, perfectly human-like androids cannot be created by humans through inorganic means, because to be able to do that, the latter must be ahead of the former in such a way, that the current hesitation may result as a childish behavior. The android will be created by a man who is not as we known him, neither just a bit cleverer, nor a bit more hairless: the cyborg.

According to Donna Haraway (Haraway, 1991), a cyborg is “(...) a *cybernetic organism, a hybrid of machine and organism, a creature of social reality as well as a creature of fiction*”. Ludwig Wittgenstein asserted that even if science were to comprehend all secrets of the universe, the essential problems of humanity would not be even remotely touched (Wittgenstein, 1922). It can be easily deduced from this that the same process of “removing” carried out by studies on artificial intelligence on humans could be understood in a similar way for science in general on nature.

Once humanity has transcended its nature through the incorporated machine, created starships, space colonization technologies, and above all, after successfully creating an android, what would be the next step? How would the inherent “forward!” of science and thought manifest?

It seems obvious that the answer to these questions could at least find hope in the physical transfer to another planet, for monitoring the android left to its own history. On the other hand, what would result from observing that the history of the android would unfold over centuries in a manner like that of humans, I dare not even imagine, but posthumans might, and it will be up to them.

Just as living beings had the courage to free themselves from water and earth, just as humans detached their hands from the duties of locomotion to dedicate them to other purposes (Leroi-Gourhan, 1993),

in the same way, humans will leave their bodies without dying to experience a new conception of freedom, experience, as well as self-limitation. To do this, they must finally relinquish the sense of touch, obliterate egotism, and ultimately abolish the only remaining barrier between themselves and the new Nature they have created. He must become homo *absolutus*, the supremely detached man. If “I” stands for “Intelligence”, then this is the “I²”.

7. Conclusions and Paths for the Future

By reaching the extreme consequences of some approaches to ethical concerns in Artificial Intelligence, we have shown that a clear and purposeful research direction is still to be found. Currently, almost all principles are accepted and taken as starting points to build up methods and solutions, which cannot be held as effective for very long, despite their formal correctness. Working with data is the principle of science; nonetheless, in this specific case, an imaginative effort is required to accept as data events that have not yet happened but are most probably approaching. This allows us to be ahead of what expects us and build more effective strategies.

In general, ethical concerns should be explicated by the traditional means of philosophy and art and addressed in public debate, which can better interpret and direct human behavior, building awareness. On the contrary, the strictness of law would just compel bad-intentioned scientists to find new and more creative (but most of all, perfectly lawful) ways to escape them.

Besides dealing with ethical challenges posed by current AI (such as algorithmic bias), it is argued that the time is ripe to deal with imminent developments of AI (AGI) not only from an ethical but even more from an existential perspective. Existing tools serving this purpose can be sharpened even by leveraging AI itself. Lightning

technical improvements in human history have never raised such a great danger for human existence.

The improvement of systematization and automation of Futurology by means of Natural Language Processing includes:

(a) Approaching the prioritization task of Horizon Scanning by leveraging Topic Modeling as an NLP tool for identifying pivotal trends.

(b) Handling Future Wheels as Knowledge Graphs to unlock automated Knowledge Expansion and Integration (labeling relations as the first step towards this goal).¹⁴

(c) Harvesting textual data featuring discussions of past new trends aiming at performing Sentiment Analysis and comparing predictions with actual realizations.

At the same time, measures to ensure responsible AI development and deployment should focus on three key areas:

(a) Introduction of AI-related “Millennium Problems,” “Dos and Don'ts,” Guidelines, and Binding Rules for research at both academic and industrial levels.

(b) Popularization of novel technologies by fostering a virtuous adaptation in society (prizes, scholarships, and grants for projects

¹⁴ Raw data can show us insights that we are not ready or willing to see. For this reason, automating part of the analysis and insights generation can represent a game changer for our perception of the reality that is unfolding in front of our eyes.

leveraging AI to stimulate curiosity, creativity, philanthropy, awareness, knowledge, and wisdom).

(c) Involvement of experts from other fields in the ongoing discussion in AI ethics, such as Deontic Logics, Philosophy of Science, Philosophy of Mind, Futurology, as well as spiritual leaders.

Acknowledgments

This work has benefited from the support of my supervisors, Dr. Maria Pia Di Buono and Prof. Dr. Johanna Monti from the UniorNLP Lab at the University of Naples "L'Orientale," and the kind endorsement of the Department of Digital Contemporary History (C²DH) of the University of Luxembourg.

References

- Abbott, R. (2016) "I Think, Therefore I Invent: Creative Computers and the Future of Patent Law", *Boston College Law Review*, 57(4), pp. 1079-1126.
- Acquisti, A., Taylor, C. and Wagman, L. (2016) "The economics of privacy", *Journal of Economic Literature*, 54(2), pp. 442-492.
- Amodei, D. et al. (2016) "Concrete Problems in AI Safety", *arXiv preprint arXiv:1606.06565*.
- Arkin, R.C., Ulam, P. and Wagner, A.R. (2012) "Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception", *Proceedings of the IEEE*, 100(3), pp. 571-589.
- Awad, E. et al. (2018) "The Moral Machine experiment", *Nature*, 563(7729), pp. 59-64.

- Benjamin, W. (2012) *Aura e Choc-Saggi sulla teoria die media*. A. Pinotti and A. Somaini. (eds.), Piccola Biblioteca Einaudi.
- Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bossmann, J. (2016) "Top 9 ethical issues in artificial intelligence", *World Economic Forum*, 21 October. Available at: <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/> (Accessed: 15 February 2025).
- Britannica (2024) "Human intelligence psychology", *Encyclopedia Britannica*. Available at: <https://www.britannica.com/science/human-intelligence-psychology> (Accessed: 15 February 2025).
- Bryson, J.J. (2018) "Patience is not a virtue: the design of intelligent systems and systems of ethics", *Ethics and Information Technology*, 20(1), pp. 15-26.
- Darling, K. (2020) "2020: The Year of Robot Rights", *MIT Press Reader*, 12 December. Available at: <https://thereader.mitpress.mit.edu/2020-the-year-of-robot-rights/> (Accessed: 15 February 2025).
- diagrams.net (2024) *draw.io - Diagrams for Everyone*. Available at: <https://www.drawio.com/> (Accessed: 15 February 2024).
- European Commission (2024) *Liability rules for Artificial Intelligence*. Available at: https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en (Accessed: 15 February 2025).
- European Parliament and Council (2024) „Regulation (EU) 2024/1689 of the European Parliament and of the Council on harmonised rules for artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts“, *Official Journal of the European Union*. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689> (Accessed: 15 February 2025).

- Frey, C.B. and Osborne, M.A. (2017) "The future of employment: How susceptible are jobs to computerisation?", *Technological Forecasting and Social Change*, 114, pp. 254-280.
- Gabriel, I. (2020) "Artificial Intelligence, Values, and Alignment", *Minds and Machines*, 30(3), pp. 411-437.
- Gesley, J. et al. (2019) *Regulation of Artificial Intelligence in Selected Jurisdictions*. Washington: The Law Library of Congress, Global Legal Research Directorate.
- Glenn, J.C. (1972) "Futurizing Teaching vs Futures Course", *Social Science Record*, 9(3).
- Graham, C. (2024) "Using ChatGPT for Foresight: Futures Wheel", *Medium*, 10 January. Available at: https://medium.com/@christian.graham_49279/using-chatgpt-for-foresight-futures-wheel-8e79eecfe86b (Accessed: 15 February 2025).
- Haraway, D. (1991) "A Cyborg Manifesto: Science, Technology, and Socialist Feminism in the Late Twentieth Century", in Simians (ed.) *Cyborgs and Women: The Reinvention of Nature*. New York: Routledge, p. 149.
- IBM (2024) *Sentiment Analysis*. Available at: <https://www.ibm.com/topics/sentiment-analysis> (Accessed: 28 February 2024).
- Kelly, M. (2024) "EU retreats on AI liability rules amid US competition concerns", *Financial Times*, 13 February. Available at: <https://www.ft.com/content/3e4e7986-b4d2-4d98-940d-55b57d863fcd> (Accessed: 15 February 2025).
- Leroi-Gourhan, A. (1993) *Gesture and Speech*. Anna Bostock Berger (trans.). Cambridge: MIT Press.
- Macciola, A. (2019) "Bad, biased, and unethical uses of AI", The Enterprisers Project, 29 August. Available at:

- <https://enterprisersproject.com/article/2019/8/4-unethical-uses-ai> (Accessed: 15 February 2025).
- Matthias, A. (2004) "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology*, 6(3), pp. 175-183.
- Mermelstein, J.J. (2022) "The Evolutionary Roots of Anxiety and its Implications for Socialization and Group Cohesion", *Journal of Psychiatry and Psychiatric disorders*, 6, pp. 243-248.
- Ministry of Health Singapore (2021) *Horizon Scanning Methods and Process Guide*, Version 1.0. Singapore: Ministry of Health.
- Musk, E. (2019) "Neuralink: An Integrated Brain-Machine Interface Platform With Thousands of Channels", *Journal of Medical Internet Research Publications*. doi: 10.2196/16194.
- O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
- Ord, T. (2020) *The Precipice: Existential Risk and the Future of Humanity*. London: Hachette Books.
- Powell, A. (2020) "AI Revolution in Medicine", *The Harvard Gazette*, November. Available at: <https://news.harvard.edu/gazette/story/2020/11/risks-and-benefits-of-an-ai-revolution-in-medicine/> (Accessed: 15 February 2025).
- Raghavachary, S. (2020) "Intelligence - Consider This and Respond!" In Samsonovich, A.V., Gudwin, R.R. and Simões, A.S. (eds.) *Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICAAI 2020. Advances in Intelligent Systems and Computing*, vol. 1310. Springer, p. 48.
- Ramamurthy, U. et al. (2006) "LIDA: A working model of cognition" In *Proceedings of the 7th International Conference on Cognitive Modeling*.
- Roberts, F. (2024) "Commission plans to withdraw AI liability directive draw mixed reactions", *EURACTIV*, 13 February. Available at: <https://www.euractiv.com/section/tech/news/commission-plans-to->

- withdraw-ai-liability-directive-draw-mixed-reactions/ (Accessed: 15 February 2025).
- Russell, S. (2022) "Artificial Intelligence and the Problem of Control" In Werthner, H. et al. (eds.) *Perspectives on Digital Humanism*. Cham: Springer.
- Schneider, S. (2019) *Artificial You: AI and the Future of Your Mind*. Princeton: Princeton University Press.
- Searle, J.R. (1980) "Minds, brains, and programs", *Behavioral and Brain Sciences*, 3(3), pp. 417-424.
- Sun, R. (2007) "The importance of cognitive architectures: An analysis based on CLARION", *Journal of Experimental & Theoretical Artificial Intelligence*, 19, pp. 159-193.
- Sun, R. (2020) "Potential of full human-machine symbiosis through truly intelligent cognitive systems", *AI & SOCIETY*, 35.
- UNESCO (2021) *Recommendation on the Ethics of Artificial Intelligence*. Available at: <https://en.unesco.org/artificial-intelligence/ethics#recommendation> (Accessed: 15 February 2025).
- Wallach, W., Franklin, S. and Allen, C. (2010) "A conceptual and computational model of moral decision making in human and artificial agents", *Topics in Cognitive Science*, 2, pp. 454-485.
- Willings, A. (2021) "27 ways military tech changed our lives", *Pocket-lint*, 19 May. Available at: <https://www.pocket-lint.com/gadgets/news/143526-how-military-tech-changed-our-lives> (Accessed: 15 February 2025).
- Wittgenstein, L. (1922) *Tractatus logico-philosophicus*. London: Routledge & Kegan Paul.

Chapter 3

Artificial Intelligence and Reflective Equilibrium

Denis Coitinho¹

The aim of this chapter is to reflect on the need for moral standards for driving autonomous vehicles (AVs), and to propose a process of reflective equilibrium (RE) for achieving this. As such, I wish to begin with an investigation into moral disagreement in order to decide how we should proceed in cases of uncertainty, and I will argue that we should adopt a procedure which brings together a range of normative criteria. Following this, I will set out the approach of collective reflective equilibrium in practice (CREP) proposed by Savulescu et al. (2021), which corrects the results of the *Moral Machine* experiment, and describes the principles of a general policy for regulating the use of AVs. My next step will be consider the limits of CREP, in the sense that it omits the normative criteria of virtues and does not form a coherent system of beliefs which is sufficiently wide-ranging. I will end the paper by proposing the use of a collective and wide reflective equilibrium (CWRE) which takes into account the normative plurality of a democratic society, whilst describing a methodology for dealing with this issue.

¹ Denis Coitinho is Full Professor of Philosophy at University of Vale do Rio dos Sinos (Brazil) and visiting scholar at University of Oxford (UK) and Harvard University (USA).

Keywords: Artificial intelligence, autonomous vehicles, moral disagreement, normativity, reflective equilibrium.

1. Introduction

The technology related to artificial intelligence (AI)² is already a constant feature of our daily lives, in terms of the recommendations of songs and movies we receive through streaming services, such as Spotify and Netflix, the e-mail filters we use, or navigational apps, such as Waze. These systems are not seen as a problem since they have the effect of making our lives easier. However, this technology is progressively expanding into areas where its impact will be even greater, such as taking decisions in situations involving risk, establishing priorities and making complex judgements, all of which involve moral elements.

In the light of this, it is important to analyse the algorithms which govern this technology. For example, autonomous vehicles need to take decisions concerning the distribution of risk between passengers, pedestrians and cyclists. Algorithms which are already in use in the health and justice systems of some countries can establish priorities in relation to who should receive a heart transplant, or who should be granted parole or given a longer prison sentence (Bonnefon, Sharif and Rahwan, 2020, pp. 109-111).

The aim of AI³ may be to simplify and add quality to our daily lives by improving human decisions, avoiding road accidents, saving lives

² AI refers essentially to the use of hardware and software to carry out tasks which usually require the use of intelligence when accomplished by human beings. See Frankenfield, 2022, and Copeland, 2001.

³ There are two types of AI. Strong AI involves software which seeks to reason and make decisions in the same way that human beings do. In other words, it attempts to reproduce the processes which human minds go through when they

during armed conflicts, optimising the process of organ donations and transplants, or even preventing violent crimes. In theory, AI algorithms can make decisions soundlessly and without cognitive bias, in a logical and rational way, whilst avoiding any kind of partiality or prejudice. However, this does not always work out as expected, since many decisions will inevitably involve ethical standards and complex moral evaluations. For example, should autonomous vehicles always endeavour to avoid causing injury to passers-by, even if this entails sacrificing the safety of the vehicles' occupants? Should lethal automatic weapons always have the aim of winning conflicts even when this entails killing innocent civilians? Should children always be given priority for organ transplants even when an older patient is genetically more suitable to receive the organ in question? Should the algorithms used in legal cases always endeavour to reduce re-offending even when this results in unfair discrimination against black offenders, as tends to happen with the COMPAS programme used in the US legal system?⁴

In the case of autonomous vehicles (AVs)⁵, which are expected to cause less stress and offer better road safety, several problems have

deliberate and decide. Weak AI, on the other hand, simply aims to provide intelligent assistance to human agents, and the only proviso is that machines should be better at taking certain decisions than their human counterparts in an efficient way and within parameters defined by human beings themselves. See Rozenfeld, 2016; Etzioni and Etzioni, 2017, pp. 410-411.

⁴ Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a programme used by 48 states in the US judicial system to predict which criminals will re-offend, thus influencing decisions concerning parole, bail and sentencing. However, a test carried out by *ProPublica* has shown that COMPAS is often subject to error, as well as being biased against black people, in that it tends to identify them as future criminals and places them in the category of potential repeat offenders twice as often as white people. See Larson et al., 2016; Mesa, 2021.

⁵ AVs are complex systems which make use of *deep learning* and can be categorised as a form of strong AI, as is the case with the Tesla and Waymo

already been noted, making it a matter of urgency to discuss whether the vehicles' AI should be programmed with certain moral standards. In 2018, an Uber vehicle which was undergoing tests collided with Elaine Herzberg when she was riding a bicycle in Tempe, Arizona. Those investigating the incident said that the test driver, Rafael Vazquez, was watching a television programme at the time, and was therefore accused of manslaughter. It appears, however, that the vehicle's automatic system wrongly identified Ms. Herzberg and her bicycle as an imminent danger of collision with the vehicle itself. In 2019, a driverless Tesla vehicle in Texas collided with another vehicle and caused the death of two people. According to the local authorities, the driverless vehicle was travelling at high speed and without a test-driver. Despite these incidents, some of the largest automobile manufacturers have announced plans to introduce completely autonomous vehicles by 2024, and in fact there are already a significant number of driverless AVs on the roads today. Examples of this are the Baidu robotaxis which have been circulating in the Chinese city of Wuhan since December 2022, and the Waymo and Cruise companies which have been providing robotaxi services in Phoenix, Arizona, and San Francisco, California.⁶

Given the damage which can clearly be caused by driverless vehicles, it is ethically vital for AVs to be able to distinguish between correct and incorrect decisions, that is, they need to have a capacity for moral

vehicles. They are programmed to collect and process information, draw conclusions, and change their behaviour accordingly, without any human intervention or guidance. See Kaur and Rampersad, 2018, pp. 87-88.

⁶ For more information regarding these cases, please see the following news items: Uber's self-driving operator charged over fatal crash (BBC News, 16/09/2020); Two killed in driverless Tesla crash, officials say (The New York Times, 18/04/2021); China's Baidu operates taxi night in Wuhan (English News, 27/12/2022); Waymo expand service area in 2 cities (The Robot Report, 16/12/2022).

reasoning (Wallach and Allen, 2009, pp. 3-12). In a similar manner, many AI researchers affirm that, if these vehicles can make thousands of cognitive decisions based on the information they receive, they should also be capable of making ethical decisions. As Anderson and Anderson have stated, "Ideally, we would like to be able to trust autonomous machines to make ethical decisions on their own, and this requires that we create an ethic for these machines" (Anderson and Anderson, 2011, p. 1). This obviously shows that AVs using AI need to be programmed with ethical orientation, since they have to make decisions which do not involve human intervention but which can be damaging to all the parties involved (Anderson and Anderson, 2011, pp. 1-4).

This ethical requirement is necessary because there may be cases in which an AV may have to choose between negative alternatives. For example, it may have to decide between avoiding a pedestrian whose life is in danger or saving the lives of its passengers. Other examples would be to give priority to children instead of old people or to women rather than men; or to decide whether it would be correct to knock down a non-human animal in order to avoid colliding with a roadside barrier and causing material damage to the vehicle while saving the lives of its passengers. It is worth mentioning here that an AV without passengers, such as a robotaxi, has a moral obligation to avoid colliding with both human and non-human animals, even if this means crashing into a lamppost, for example. Although this would cause material damage to the vehicle, it would clearly save lives. Another possible scenario would be for the AV to swerve in order to avoid knocking down and killing a pedestrian, even if this meant crashing into a roadside barrier and injuring its passengers. These moral considerations are obviously important because they reflect on certain essential human values, such as dignity, the value of life, and

equality. The question is, who should make decisions concerning this issue?

In the light of the examples given above, we can see that the use of AI in AVs raises very significant questions which demand reflection on the capacity and nature of human and artificial agency, together with moral and legal responsibility, blame and punishment. Whilst recognising the vital importance of these considerations, this article will nevertheless continue to focus on the normative standard which can be used to guide AI decision-making in driverless AVs. The main problem is that, although there is general agreement that these vehicles should be able to use moral standards for making decisions, there is general disagreement about what makes certain actions right or wrong. This is because moral disagreement is an easily identifiable social phenomenon, in the sense that we also disagree about which moral theory should be chosen in making these decisions. For example, should we use the principle of maximisation of well-being which is typical of utilitarian theories, or should we use the deontological principle of universalisability? Or alternatively, should we refer to the virtue ethics for guidance in this matter? In addition, there is disagreement as to whether we should use any moral theory at all in cases of this nature. Some experts affirm that AVs should simply follow existing legislation, such as the highway, penal and civil codes, without any recourse to moral theories (Etzioni and Etzioni, 2017, p. 415). In the light of all this, I believe that there should be a more detailed investigation into moral reasoning and that we should first reflect on the complexity of normativity in order to identify the boundaries of traditional moral standards. With this in mind, the strategy I will adopt is to propose a specific type of RE as the basis of a moral reasoning for identifying the normative standards which can guide the AI of AVs.

With this in mind, I show the order in which I will deal with topics in the rest of this paper. I will begin with an investigation of moral reasoning to determine how we should make decisions on cases of uncertainty, and I will argue that we should make use of a process which brings together a range of normative criteria. I will then describe the procedure concerning collective reflective equilibrium in practice (CREP) as proposed by Savulescu, Gyngell and Kahane (2021), which corrects the results of the *Moral Machine* experiment (Awad et al., 2018) and outlines the normative principles of a general policy for regulating AVs. My next step will be to identify the limits of CREP, in the sense that it omits the normative criteria of virtues and does not provide a coherent system of beliefs which is sufficiently wide-ranging. Finally, I will propose the use of a collective and wide reflective equilibrium (CWRE) as a way of including the normative plurality which is fundamental to a democratic society, and outline a methodology for dealing with this issue.

2. Moral Disagreement

One of the main problems with the development and application of moral standards in AVs is the high level of open disagreement concerning whether a certain action is right or wrong. In a similar way, we may disagree about what constitutes a good life, or which normative moral standard is the best basis for the ethical decisions we make, and questions about abortion, euthanasia and animal rights are good examples of the phenomenon of moral disagreement. Is the correct action one which maximises the well-being of those involved, or is it one which is universal and does not treat the human agents as a mere means? Should we always keep our promises, even when this leads to an undesirable outcome, or would it be right to lie in order to save the life of an innocent person who is being persecuted? We may

even disagree about the meaning of justice itself. Is it freedom or equality, or is it a combination of both these values?

We may also disagree about which moral theory should guide our ethical decisions. Is utilitarianism better than a deontological model such as Kantism because the former takes into account the best results of a certain action, or is Kantism superior to the principle of maximising well-being, since the latter concept may lead to the violation of individual human rights? Other ethicists may affirm that neither utilitarianism nor Kantism are the best available moral theories, since they are principalistic models which are not connected to the emotional and dispositional structure of human agents. In that case, it would seem better to choose the virtue ethics as a suitable moral theory, where the correct action must be approved by a virtuous agent, because this approach places the question of how we should live our lives above that of how we should act.

As regards normative disagreement, it is important to recognise that there are significant theoretical disputes about whether we should make decisions based on moral principles or on our own intuitions concerning what is right or wrong, especially when we are discussing the need for autonomous machines to be programmed with ethical standards, in the sense that they are considered to be artificial moral agents which are capable of moral reasoning. Many people affirm that to effect moral reasoning we need to apply certain ethical principles in each specific case. For example, in order to know whether we should approve voluntary euthanasia, would it be sufficient to apply the principle of maximisation to see if such an action will improve the well-being of those involved. Alternatively, would it be sufficient to apply the principle of universalisability and non-instrumentalisation in order to find out whether euthanasia should be approved by a universal rule which does not consider the agent merely as a means

towards an end. Many ethicists affirm that principlism is limited by its abstraction and rigour, and therefore we should adopt a perspective which is based on agents' intuitions and decisions, whilst at the same time seeking a means for improvement on this basis, such as the virtue ethics, which would take into consideration the feelings and experience of the agents. This dispute is known in the relevant literature as the confrontation between the *top-down* and *bottom-up* moral models.⁷

A more fruitful way of dealing with the problem of AVs in relation to moral uncertainty may be to adopt a pluralist perspective. Instead of having to choose one of the traditionalist theories available, that is, the utilitarian or deontological theories, or the virtue ethics, it might be possible to unite all of these in a first-order normative theory. We would thus have at our disposal a hybrid moral theory which makes use of a variety of distinct normative criteria which are more coherent with each other. This is because exclusive moral normativity seems to be unsatisfactory for dealing with complex problems such as those relating to AI, which often demand a range of normative criteria for identifying solutions.

A traditional moral theory uses, in general, only one normative criterion, such as the maximisation of well-being in utilitarianism, universalisability and non-instrumentalisation in Kantism, or even a virtue agent such as prudence in the virtue ethics. In many situations, other normative criteria are needed to solve specific problems, which

⁷ This dispute is very common in the field of AI ethics, since ethics experts and AI scientists are divided in relation to the best model to be adopted with the aim of developing artificial moral agents who can make decisions or, most radically, establish moral reasoning. For example, Wallach and Allen (2009, pp. 117-124) affirm that we should use a model which contains both *top-down* and *bottom-up* approaches, and for them the virtue ethics is the most appropriate one for this particular task.

highlights difficulties concerning the internal coherence of a particular moral theory. For example, how can a utilitarianist theory make use of a deontological criterion of freedom without being incoherent? How can a deontological theory make use of a criterion of maximisation of well-being without manifesting internal contradiction? And how can a virtue ethics make use of a principle of human dignity in order to ensure human rights without compromising its own structure? Even rule utilitarianism, which seeks to combine the principle of maximisation of well-being with a universal rule leading to the great good must face the problem of internal coherence.⁸

In order to further clarify the limitation outlined above, it is useful to refer to what are known as *trolley problems*.⁹ In general, these problems are concerned with presenting moral dilemmas to specify which intuitions are manifested by the agents involved, with the aim of identifying the corresponding moral principle. One of the most widely known dilemmas is related to the question of sacrificing a single person in order to save the lives of others. For example, when people are asked what they would do if a runaway train travelling at 100 kms per hour was about to knock down five workers on the main line, but by pulling a lever the train could be diverted on to a branch

⁸ The main objection to rule utilitarianism is that it would be incoherent, since the theory permits actions which do not maximise well-being, even though it is tacitly linked with this maximisation and therefore would no longer be considered as utilitarianist. See Raizci, 2016, pp. 857-860.

⁹ *Trolley problems* are a series of mental experiments in ethics and psychology involving a range of ethical dilemmas relating to whether it is right to sacrifice a single person in order to save a larger number of people. The main point here is that we need to make use of specific cases in order to reach certain ethical principles. Originally, the dilemma identified by Foot was from the point of view of the train driver and not of a bystander. See Foot, 1967; Thomson, 1976 and 1985.

line with only one worker on it, most people would choose the second option.

However, if there was an option to push a healthy person from a road bridge in order to save the five workers on the main line, most people would not agree with this. The question is, why is the principle which appears to be correct in the first case considered to be wrong in the second? The principle in the first case clearly indicates that it is correct to sacrifice one person in order to save five; in other words, we should try to save the highest number of people possible because of the preciousness of human life, and this is a good example of the principle of the maximisation of human well-being. On the other hand, the principle in the second case is that it is wrong to kill even a single human being whatever the consequences may be, which places it in direct opposition to the principle described in the first case.

This is similar to the situation in which it would be considered wrong to kill a healthy person in order to transplant their organs into five gravely ill people who urgently need this treatment to survive. The conflict of principles here is clear; on the one hand we know that it is right to try and save as many lives as possible, while on the other hand it is wrong to kill even a single human being.

In the context of the aforementioned *trolley problems*, utilitarianism would assert that it is right to save the greatest number of lives possible, since its fundamental moral principle is that of the maximisation of well-being, and this is why pulling the branch line lever to divert the train and save the lives of the five workers is considered to be the correct action, even though it may entail the death of one other worker. This means that we must accept the principle of sacrifice for the general good. In accordance with this principle, it would also be considered correct to push a healthy person off the road

bridge in order to save the five workers, or to carry out the transplant described above in order to save the lives of five gravely ill people.

The problem is that in line with our common moral senses it is still wrong to take the life of a single human being, even when the consequences of this may be positive. Kantism, on the other hand, would assert that it is wrong to divert the train in order to save the lives of the five workers, since the principle of non-instrumentalisation prohibits any action which would consider a single person to be a simple means towards protecting the well-being of others. This seems problematic because we ourselves agree that we must save the greatest number of people possible because we believe in the value of human dignity. And even if both utilitarianism and Kantism also wished to value these common moral judgments, they would have difficulties with the internal coherence in the theory itself.

In the light of the above, it would be very convenient if there was a hybrid moral theory which contained a variety of normative features, such as the utilitarianist criterion of best consequences, the deontological criterion of non-instrumentalisation (which guarantees universal respect for human dignity), and even a set of virtues which constitute normative standards (which are not based on right action but on good life). This would help us to achieve a greater understanding of right and wrong by means of coherent alignment with this plurinormative set of values. And if that is not yet available to us, then I think we can at least try to use a strategy or methodology to connect these various normative criteria coherently.

In this paper I intend to test the scope of the application of the RE method in identifying moral standards for AVs, bearing in mind the importance of plurinormativity, since it is a procedure which is normally used to justify moral principles by means of a mutual adjustment with the considered moral judgments that one has great

confidence in, and also with factual judgments supported by certain theories that are relevant in the case, such as in justifying principles of justice through their coherence with a coherent system of beliefs, which leads to a wide reflective equilibrium (WRE).¹⁰ This appears to be a promising procedure in the context of normative-moral uncertainty precisely because it can be interpreted as a method which helps us to know what we should do in cases of doubt. It is the conclusion of a deliberative process in which reasons are considered and a certain course of action in relation to a specific problem is chosen. It can also be used as a method for underpinning discussions and justifying the making of decisions which are collective and not exclusively personal.¹¹

Keeping all of this in mind, in the remain of this chapter I shall analyse the characteristics of RE, beginning with the specific use of the method used by Savulescu, Gyngell and Kahane (CREP) in their discussion concerning the principles which should be chosen as moral standards for AVs. Following this, I shall propose a reformulation of CREP in order to incorporate the normative standard of virtues and form a wide-ranging system of coherence beliefs, thereby recognising the collective and wide reflective equilibrium (CWRE) as a pluri-normative strategy. In the end of the paper, I shall considerer that the

¹⁰ See Rawls, 1971; Daniels, 1979; Scanlon, 2003.

¹¹ Kushner et al. (1991, pp. 286-292) refer to the RE method as a systematic approach to the treatment of an ethical dilemma. In this interpretation, it is a method of decision-making which can be used to adjust existing principles and judgements in relation to a specific problem with the purpose of obtaining a justified decision. Furthermore, Bransted and Brannmark (2020, p. 357) set out a version of RE which may be considered as "(...) a tool for social reasoning about practical problems, such as how to distribute scarce resources in different domains of practice, which facilitates shared solutions". For further details concerning the RE method and its different uses in moral and political philosophy, and in epistemology, see Reznitzner (2022, pp. 1-9).

overall result of CWRE is the identification of certain parameters which will provide guidance for the operation of AVs.

3. Collective Reflective Equilibrium in Practice

In recent article, Savulescu, Gyngell and Kahane (2021) asserted that the RE method can be used to accommodate public preferences as to how AVs should react in situations of emergency, in contrast to the principles of certain ethical theories, such as utilitarianism, Kantian deontology and Rawlsian contractualism. The method proposed is called *Collective Reflective Equilibrium in Practice* (CREP), in which data on public preferences from around the world are collected and tabulated by using the online platform *Moral Machine*. Moreover, these data serve only as input for a public deliberative process which seeks to identify coherence between attitudes, behaviours and ethical principles, and is not intended to be the final word on the question, since moral intuitions may express prejudices, bias, and personal and/or corporative interests (Savulescu et al., 2021, pp. 1-2). The overall aim is to see whether considered judgments can survive when confronted with the ethical principles of utilitarianism, Kantism and contractualism. When there is inconsistency, our initial beliefs must be revised in order to arrive at a situation of adequate reflection. On the basis of this, the authors then propose a public policy to try normalise the situation (Savulescu et al., 2021, pp. 5-12). But before detailing the CREP procedure, let's briefly explain the *Moral Machine* experiment.

The *Moral Machine* experiment, carried out by researchers from MIT, consists of a survey via online platform of the preferences of people from around the world concerning how AVs should make decisions in situations of emergency where the death of someone is imminent. For example, should a car without brakes swerve to avoid colliding with five individuals at the cost of knocking down just one individual?

Should human beings be given priority over other animals? Or should AVs give priority to people who obey the highway code, or to healthy individuals over obese ones? As a result of this survey, 40 million responses in 10 languages from 233 countries were collected. After tabulating the data, nine public preferences were identified in descending order of priority: (1) Sparing humans; (2) Sparing more characters; (3) Sparing the young; (4) Sparing the lawful; (5) Sparing the higher status; (6) Sparing the fit; (7) Sparing females; (8) Sparing pedestrians; (9) Preference for the vehicle's inaction. Of these nine preferences, three stronger ones were identified: (i) Sparing human beings over other animals; (ii) Sparing more lives; (iii) Sparing young lives (Awad et al., 2018, p. 59).

In the words of the authors:

As shown in Fig. 2a, the strongest preferences are observed for sparing humans over animals, sparing more lives, and sparing young lives. Accordingly, these three preferences may be considered essential building blocks for machine ethics, or at least essential topics to be considered by policymakers. Indeed, these three preferences differ starkly in the level of controversy they are likely to raise among ethicists (Awad et al., 2018, p. 60).

For Awad et al. (2018), therefore, the public preferences collected from the *Moral Machine* experiment may serve as a model for the development of an ethics in AI machines, or as a relevant normative standard which may be taken into consideration by legislators working on an agenda for regulating AVs.¹²

¹² For a critique of this proposal, see the "The Immoral Machine", an article by John Harris (2020), in which he strongly criticises the attempts to identify a public morality for the guidance of AVs.

For Savulescu, Gyngell and Kahane (2021), it is extremely relevant to be able to have access to data on collective preferences when establishing a public policy which aims to normalise the functioning of AVs, and this is for two main reasons. Firstly, there are no moral truths available which can indicate what is right or wrong in such circumstances, especially when we consider that the most widely recognised moral theories give us different answers to these dilemmas. For example, according to utilitarianism it is correct to save more lives, since the normative criterion stipulated is that of the maximisation of well-being. However, this would be wrong in the case of Kantism, since it goes against the principle of human dignity. Secondly, in liberal democracies the legitimacy of laws depends on public support, especially in complex cases where there are high levels of disagreement. For the authors referred to above, “The challenge, then, is to find a way of integrating ethical theory and data about the spread of public intuitions while minimising the risks associated with each” (Savulescu et al., 2021, p. 5).

In addition, Savulescu, Gyngell and Kahane (2021) correctly observe that, although they are extremely relevant, the data collected may express prejudice, tribal bias, and personal and/or corporative interests. For example, they point out that the preference for sparing the young does not have the same strength in the three regions researched (Western cluster, Eastern cluster, and Southern cluster). In the Eastern cluster, the preference for saving the young is weaker than the preference for saving those who obey the highway code, whilst in the Southern cluster, which includes countries in Central and South America, the preference for saving people of higher social status is almost as strong as the preference for saving the young. The problem is that, if we are to make use of these preferences, we must decide whether we are going to choose the strongest ones globally or only those of a specific region of the world. Even more importantly, some

preferences should not be used as the basis for a public policy, since they are tribalistic in nature, and may give a higher priority to native inhabitants rather than foreigners, or to family members instead of the public at large. There may also be discrimination in terms of ethnicity, social class or gender, and since many preferences may be similarly distorted, RE should be used to correct such anomalies.¹³

Savulescu et al. (2021) therefore adopt the RE method as formulated by Rawls and adapt it in practice as a collective RE (CREP), which can be understood as a deliberative process where public intuitions and ethical theories are of equal importance for being able to identify consensus by means of adequate reflection.¹⁴ They begin with the input provided by the public intuitions, that is, the preferences collected via the *Moral Machine* survey, and filter them in order to select the considered judgments, discarding those intuitions which stem from prejudice and distortion. In CREP, the considered judgments do not use their own intuitions, and certain criteria are applied to verify the reasonability of the filtered preferences. Three preferences are then selected which may be taken as considered judgments, and these are (i), saving more lives; (ii) saving the young

¹³ Savulescu et al. (2021, p. 6) point out that the participants in the study undertaken by Awad et al. (2018) were not asked whether they would prefer to save members of their own ethnic group or nationality rather than save foreigners. They were also not asked whether they would prefer to save their own lives and those of their family rather than save the lives of the public at large. It should be mentioned here that other smaller-scale studies have also revealed similar tribalistic and self-interested characteristics.

¹⁴ In CREP, (i) Deliberators are policymakers, ethicists, and relevant medical and scientific experts; (ii) Intuitional input consists of preferences/judgments of general public/citizen juries; (iii) Theoretical input consists of ethical theories, values, principles, frameworks, guidelines and declarations; (iv) Output consists of the justified policy for the democratic process; (v) Iteration is limited. The main point here is to understand the procedure as a deliberative public process realised by a collective agent, and not as an individual deliberation realised by a single individual (Savulescu et al., 2021, p. 6).

and (iii) saving females. The next stage in the procedure is to find coherence between these considered judgments and the moral principles of the three ethical theories previously identified, namely, utilitarianism, Kantism and contractualism. The authors state that “(...) We need to check if these intuitions are actually responding to ethically plausible underlying *reasons*, which would increase our confidence in their validity. After all, even the most considered and widespread judgments can still be mistaken” (Savulescu et al., 2021, p. 8).

Kantism, for example, through the principle of human dignity, would approve the preference for saving human lives, but would disapprove the preference for saving more lives, since, according to this theory, each human life possesses the same dignity, and for the same reason it would disapprove the preference for young people and females. On the other hand, utilitarianism, on the basis of the principle of maximisation of well-being, would approve the preferences for saving human lives because of its criterion that every human life is useful. It would also approve the preferences for saving more lives and saving the young, but would reject the preference for saving females, since this does not appear to maximise well-being. In a similar way, contractualism, with the criterion of impartiality provided by the veil of ignorance, would approve the preferences for saving human lives, saving more lives, and saving the young, but would reject the preference for saving females, since under the veil of ignorance we would not know to which gender the potential victims belong.

Having completed the task of confronting public intuitions with influential ethical theories, the third step in the procedure would be to propose a justified public policy which would take as its normative criterion the preferences for saving human beings and for saving as many of them as possible. The preference for saving the young could

be left for later deliberation, since it would require undertaking empirical research. The preference for saving females would be rejected, and this would then necessitate a revision of initial moral convictions. In this way, Savulescu et al. clearly state that “the contribution of CREP is to clarify the role in this process of data about public preferences, on the one hand, and the constraints imposed by ethical theories on the other” (Savulescu et al., 2021, p. 9).

After this brief presentation of CREP, it is important to clarify why I believe that this is a promising approach. Firstly, it does not try to identify a universal moral principle which can be applied in all cases, in the sense that this principle will always be taken as true. The procedure uses a bottom-up methodology based on the intuitions shared by people from all around the world in order to test its coherence as a plural normative whole. As it is not possible to identify true moral principles in a representationalist sense, a good starting point would be to take moral judgments in which we have a high degree of confidence and test them for their coherence with ethical principles which have wide currency in society. Secondly, this strategy seems to be an interesting one because it interprets RE as a collective decision and not as the deliberation of a single individual. Since we are endeavouring to identify normative standards for AVs which will be in use all over the world, it is important that a wide range of beliefs is taken into account. In this way, the outcome of the deliberative process will be a public decision which can serve as a model for future legislation. This is also important from the point of view of political legitimacy since, in liberal democracies, public policies depend on the approval of a majority of citizens, at least at a certain level, thus avoiding the possibility that such important decisions are made only by experts or technocrats, or even by market forces. The final reason for following this strategy is that it uses a range of ethical standards to test shared moral intuitions by making

use of only the most relevant moral theories. This is important because, in the context of moral-normative uncertainty, it is more prudent to opt for plurinormativity.

Even though I may consider that CREP represents a great advance in relation to this subject, I nevertheless believe it has several limitations. For example, it does not make use of a normative criterion of virtues, something which is relevant both in the everyday language of blame and praise and in moral philosophy. In the light of this, I believe that virtue ethics should replace contractualism as the third normative-moral standard for testing shared intuitions. Furthermore, this procedure does not satisfactorily explain how shared intuitions can be filtered in order to transform them into considered judgments. We only know that this process will be executed by politicians, ethicists, lawyers, scientists, doctors and ordinary citizens, but no detail has been given regarding the filtering process itself, and this may lead to an undesirable arbitrariness. For example, in the *Moral Machine* experiment, of the nine most popular public preferences given, three were identified as having the highest priority, that is, saving human lives, saving more lives and saving the young. When CREP was used, other preferences were selected, that is, saving more lives, saving the young, and saving females, but at no point did the authors of the procedure explain how this choice was made. A third limitation is that this type of RE seems to be too narrow, since it only establishes coherence between considered judgments and ethical principles. With a wide reflective equilibrium (WRE), however, it is possible to test the initial input and ethical principles empirically, that is, with the factual judgments of certain scientific theories, thus creating a wider coherent system of beliefs.

4. Collective and Wide Reflective Equilibrium

Based on this problem outlined above, I shall now set out an alternative procedure which corrects some of the limitations of CREP, including the normative criterion of virtues and relying on the coherence of certain facts in order to establish a form of RE which is both wide and collective (CWRE). The strength of this procedure in relation to the specific case of AVs is that, even when there is a considerable degree of moral disagreement, such as that concerning which ethical theory should be selected as the normative standard, to arrive at a coherent system of beliefs between our considered moral judgments, certain ethical principles that are traditionally accepted and, in line with factual beliefs of valid scientific theories, after adequate reflection, we can form a justification for the recognition which is obtained through “(...) the mutual support of many considerations, of everything fitting together into one coherent view” (Rawls, 1971, p. 21).

The fundamental idea is that CWRE should be thought as a method for making decisions on a global scale. This procedure should be both collective and public because it is based on the shared moral intuitions which have been identified by the *Moral Machine* experiment, then filtered so that they may be taken as considered judgments, and after, they should be tested for their coherence with the ethical principles of the main theories in the moral debate, that is, maximisation of well-being, the categorical imperative, and the virtuous agent; furthermore, they should also be tested for their coherence with factual judgments of the main scientific theories. In this way, CWRE can be used as a vital tool in public reasoning, since it deals with the problem of which moral standard should guide AVs, and this is a clear sign of its plurinormativity. In addition to coherence between considered judgments, ethical principles and relevant factual beliefs,

CWRE has access to a criterion which makes it possible to give these considered judgments an initial credibility which is independent of coherence itself. The coherent system of beliefs which is thus formed should possess certain normative characteristics, such as the theoretical virtues of consistency, simplicity, accuracy and fertility. The overall aim is to achieve a result which can guide a justified public policy in a democratic process of decision-making.¹⁵

I begin, then, with the first step of CWRE, that is, identifying the input and filtering the considered judgements. Just as in CREP, it starts from the public preferences that were identified by the *Moral Machine* experiment, counting as the initial moral intuitions. As described earlier in this paper, this experiment identified nine public global preferences of which three were given priority, that is, saving human lives, saving more lives and saving the young. Also, like CREP, the deliberators are politicians, ethicists, AI scientists, lawyers, medicine doctors and other citizens. These deliberators should be specialists in the subject of AI, since their ability and experience are vital in dealing with this complex area, and they should also possess virtues such as reasonability, impartiality and open mind. So, the deliberators' next step is to analyse whether the global preferences identified by the *Moral Machine* experiment can count as considered judgments, that is, as reliable moral judgments. Differently from CREP, however, the considered judgments are not guaranteed only by the confidence of the agents concerned, but also by the epistemic and moral qualities of the deliberators themselves, thus providing an initial credibility which is independent of coherence. In line with the model proposed

¹⁵ In her recent book, Tanja Rechnitzer (2022) presents a detailed case study for the application of the RE method in order to justify the principle of precaution in cases of uncertain damage, although in her proposal the deliberation is personal and not collective. For a discussion of the characteristics of public reasoning in RE, see Brandstedt and Brannmark (2020, pp. 368-371); Baserin (2017, pp. 2-14).

by Beauchamp and Childress (2013), the deliberators should be seen as “moral judges” who possess relevant epistemic and moral virtues, such as being impartial, sympathetic and have compassion concerned the well-being of others. This is important because one of the main criticisms levelled against RE is its lack of initial credibility of beliefs. Thus, the method may be seen as having an epistemological weakness because it aims to justify a moral belief through its coherence with ethical principles and with other beliefs, but the initial beliefs are held only by something vague as the agent’s own confidence. This would imply a certain conservatism, which would be equal to depending on beliefs in the agent’s repertory which may be morally unsatisfactory.¹⁶

In order to resolve this problem, therefore, we must first look at the deliberators’ filtering of the preferences for saving human lives, saving more lives, and saving the young. They could perhaps undertake an analysis of whether any of the other six preferences culled from the *Moral Machine* experiment might also serve as considered judgments. However, since they are specialists in their own fields and are well-acquainted with the subject of AVs, in addition to possessing such virtues as reasonableness, impartiality, and open mind, I believe that it would not be controversial for them to approve the three major global preferences identified in the experiment for subsequent testing. It would in any case be necessary to justify any of the other six preferences referred in the experiment

¹⁶ Beauchamp and Childress (2013, p. 400) assert that we should begin with those moral beliefs which are part of “common morality” and are made up of “moral judgments” which possess relevant virtues such as impartiality and sympathy. Although they agree that justification is a question of coherence, they maintain that coherence alone is insufficient. It is therefore necessary, in the first place, to identify considered judgments which are acceptable without any reference to coherence. These are judgments which are reliable and in which we have confidence, and they are based on the history of moral experience rather than being the result of individual intuitions.

before adding them to the main list. In CREP, for example, the preference for saving females was identified as an initial belief and was tested for its coherence with the three ethical principles, but it failed the test in each case. But no justification was given for why a preference which was not on the final list of three and that was listed in seventh place should be identified as a considered judgment, even though the authors stated that we should accept all the public preferences as being robust, and therefore would exclude unreliable or unrepresentative data (Savulescu, 2021, p. 7).

The next step is to justify the considered judgments for their coherence with a coherent system of beliefs which comprises ethical principles and scientific beliefs. Let me start with the ethical principles. Here I will take the moral principles formulated by the three main ethical theories which are widely accepted in contemporary debate and are traditionally an important part of the history of ethics. These are the principle of maximisation of well-being (utilitarianism), the principle of universability and the principle of humanity or non-instrumentalisation (Kantism), and the principle of the virtuous agent (virtue ethics). I will look at these principles to see whether they approve the public preferences for saving human lives, saving more lives, and saving the young. This step is crucial because the theoretical reference is given by the values and principles contained in traditional moral theories, thus making it possible to contrast the considered judgments with the normative-moral criteria of the most influential ethical theories on which society relies. It is for this reason, and in contrast to CREP, that we will test the shared intuitions for their coherence with those ethical theories which have the approval and confidence of society, while choosing the virtue ethics rather than Rawlsian contractualism.

In the act utilitarianism model, the correct action is one which maximises the happiness or well-being of the persons involved, and this clearly appeals to the best consequences or the best results of the action itself. In this way, the principle of maximisation of well-being will approve the three preferences described above. The preference for saving human lives will be approved by reason of the preferential criterion of the usefulness of life, and this criterion will also approve the preferences for saving more lives and saving the young, since the latter will obviously have a longer span of utility in the future than older people. In the Kantian model, in turn, the correct action is one that would be approved by a rule which one wants to be universalized and that does not instrumentalise people. Thus, the principle of humanity or non-instrumentalisation will approve the preference for saving human lives, but will disapprove the preference for saving more lives, in the sense that all human lives possess the same dignity, and, for the same reason, it will disapprove the preference for saving the young.¹⁷

Finally, in virtues ethics, the correct action is one which will be approved by a virtuous agent, who deliberates in the best way to achieve a successful end by identifying an appropriate mean.¹⁸ In this model, the intellectual virtue of prudence is key, since the aim is to guarantee the most adequate means for the realisation of a good end. In addition to prudence, however, the agent should be able to rely on the moral virtues which will allow him to identify the mean between

¹⁷ For a description of utilitarianism and Kantism, see Brink (2006, pp. 380-4230) and Hill (2006, pp. 480-514).

¹⁸ Here I follow the proposal made by Hursthouse (1997, pp. 225-226) for a model of virtue ethics which is similar to the consequentialist and deontological models, and is explained as follows: "P.1. An action is right if it is what a virtuous agent would do in the circumstances. P.1a. A virtuous agent is one who acts virtuously, that is, one who has and exercises the virtues. P.2. A virtue is a character trait a human being needs to flourish or live well".

the two extremes in each case. In this specific case, I believe that the agent should be able to rely on benevolence in order to consider the well-being of agents, and, specially, justice so that each person can receive his due impartially. In this way, the virtuous agent will approve the preferences for saving human lives and saving more lives by reason of the idea that virtue is a disposition of character which makes it possible to live a good, flourishing life, something which is strongly associated with the condition of rationality. However, it is not certain how the virtuous agent would judge the preference for saving the young, even though there are good reasons for this in terms of the idea of caring for the most vulnerable. On the other hand, a prudent agent would want to investigate this preference more thoroughly, by looking at the existing legislation and what negative consequences the case might have for the older members of society.

The final stage of this procedure is to contrast the coherence of considered judgments and ethical principles with the factual judgments which have been made by theories which are widely accepted and are relevant to the subject in hand, and in contrast with legislation. The overall aim of this approach is to achieve a wider-ranging coherent system of beliefs. This is different from CREP, which proposes a narrow RE (NRE), whereas my proposal is for a WRE which connects values with facts. In relation to this, it is possible to analyse how both national and international laws would assess these three preferences, as well it could take into account the reports and regulations of ethics committees that have addressed the issues involved with AI.¹⁹ An example of this is the report of the ethics commission of the German Ministry of Transport and Digital Infrastructure, which has dealt with the specific problems concerning AVs, and has expressly prohibited any distinctions based on

¹⁹ For the normative content of the various existing regulations governing AI, see Floridi, 2021, pp. 5-17

individual human characteristics, such as age, gender, and physical or mental capacity.²⁰ In addition, we may consider the parameters utilised in global medicine concerning the distribution of goods which are in short supply. For example, during the COVID-19 pandemic health professionals made use of the criterion of the maximisation of well-being to give priority to young and healthy individuals with fewer co-morbidities in deciding who should be admitted to medical ICUs. This approach did, in fact, save more lives, even though it did not sit well with moral intuitions about saving the lives of the sickest and most vulnerable members of society.²¹ Furthermore, the deliberators also could consider economic factors relating to public welfare.

5. Conclusion

The overall result of CWRE is the identification of certain parameters which will provide guidance for the operation of AVs. One of these parameters is the priority for saving human lives rather than those of other animals. It may also be that the preferences for saving more lives and saving the lives of the young should be a key part of this proposal, but this will depend on the job of the real deliberators. The fundamental thesis is that all the parameters adopted must be justified both in terms of the context of their construction and of their acceptance in public way. They must be based on decisions which are both collective and democratic, since they involve the security of all members of the population. Public participation in these decisions is therefore of paramount importance, as it will serve as an essential

²⁰ See the report entitled Ethics Commission. *Automated and connective driving*, 2017. <https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf>.

²¹ On this subject, see Grover et al., 2020, pp. 889-901.

antidote to a purely technocratic approach and guided solely by market forces.²²

Last of all, it is important to emphasise that considered judgments, ethical principles and factual judgments together form a system of beliefs which appear to be coherent, whilst at the same time possessing other theoretical virtues, such as simplicity, accuracy and fertility, by reason of their fundamental attributes. One of the main characteristics of the system constituted is its plurinormativity, which includes the principle of the maximisation of well-being, the principle of human dignity, and the virtues of prudence, benevolence and justice as a coherent whole. Another vital characteristic is the system's connections between values and facts, using scientific beliefs and legislation as a means of testing the reasonability of considered judgments and ethical principles. A final characteristic of the system I would like to draw attention to is its democratic dimension, since it is based on public deliberation which relies on the acceptance of all the parties involved.

I therefore believe that CWRE offers a suitable solution to the problem under discussion in this paper. It will, of course, be necessary to investigate the properties of the system in more detail and compare it with other alternative systems. But given our normative-moral disagreement, it would seem that adopting a policy which is more coherent with considered judgments, with trusted ethical models, and

²² John Tasioulas (2022, pp. 235-239) has produced an interesting ethical treatment of AI which transcends the limitations and distortions of the utilitarian model of the maximisation of well-being, and in which he proposes a paradigm of humanist ethics which defends the pluralism of values and the justice of procedures. He states that we should not simply focus on results and on individual and collective participation in the construction of rules for the guidance of AI.

with relevant factual data, signifies a high degree of justification within a democratic society.

References

- Anderson, M. and Anderson, S. (2011) "General Introduction" In: Anderson, M.; Anderson, S. L. (eds.) *Machine Ethics*. New York: Cambridge University Press, 1-4.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, JF., and Rahwan, I. (2018) *The Moral Machine Experiment*. *Nature*, 563: 59-64.
- BBC News (2020) "Uber's Self-Driving Operator Charged Over Fatal Crash". *BBC News*. September 16.
<https://www.bbc.co.uk/news/technology-54175359>.
- Beauchamp, T. and Childress, J. (2013) *Principles Of Bioethical Ethics*. Oxford: Oxford University Press.
- Bonnefon, JF., Sharif, A. and Rahwan, I. (2020) "The Moral Psychology of AI and Ethical Opt-Out Problem" In: Liao, Matthew (ed.) *Ethics And Artificial Intelligence*. New York: Oxford University Press, 109-126.
- Brandstedt, E. and Brännmarck, J. (2020) "Rawlsian Constructivism: A Practical Guide To Reflective Equilibrium", *The Journal of Ethics*, 24: 355-373.
- Brink, D. (2006) "Some Forms and Limits of Consequentialism" In: Copp, David (ed.) *The Oxford Handbook of Ethical Theory*. New York: Oxford University Press, 380-423.

- Copeland, J. (2001) *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.
- Daniels, N. (1979) "Wide Reflective Equilibrium and Theory Acceptance In Ethics", *The Journal of Philosophy*, 76(5): 256-282.
- English News (2022) "China's Baidu Operates Taxi Night In Wuhan", *English News*. December 27.
<https://English.News.Cn/20221227/C06149e517884fabb79d1b0cad7950d1/C.Html>.
- Ethics Commission (2017) *Automated And Connected Driving*. Federal Ministry Of Transport And Digital Infrastructure. <https://www.bmvi.de/SharedDocs/En/Publications/Report-Ethics-Commission.Pdf>.
- Etzioni, A. and Etzioni, O. (2017) "Incorporating Ethics Into Artificial Intelligence", *The Journal of Ethics*, 21(4): 403-418.
- Floridi, Luciano; Cows, Josh. A Unified Framework Of Five Principles For Ai In Society. In: Floridi, L. (Ed.). *Ethics, Governance And Policies In Artificial Intelligence*. Berlin: Springer, 2021, 5-17.
- Foot, P. (1967) "The Problem of Abortion and The Doctrine of The Double Effect", *Oxford Reviews*, Vol. 5. Reimp. In: Foot, P. 1978. *Virtues And Vices*. Oxford: Blackwell.
- Frankenfield, J. (2022) "Artificial Intelligence: What it is And How it is Used", *Ivestopedia*. <https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp>.
- Grover, S., McClelland, A. and Furnham, A. (2020) "Preferences for Scarce Medical Resource Allocation: Differences between Experts and the General Public and Implications for The Covid-19 Pandemic", *British Journal of Health Psychology*, 25: 889-901.
- Harris, J. (2020) "The Immoral Machine", *Cambridge Quarterly of Healthcare Ethics*, 29: 71-79.

- Hill, T. (2006) "Kantian Normative Ethics" In: Copp, David (ed.) *The Oxford Handbook of Ethical Theory*. New York: Oxford University Press, 480-514.
- Hursthouse, R. (1991) "Virtue Theory And Abortion", *Philosophy & Public Affairs*, 20(3): 223-246.
- Kaur, K. and Rampersad, G. (2018) "Trust in Driverless Cars: Investigating Key Factors Influencing the Adoption of Driverless Cars", *Journal Of Engineering and Technology Management*, 48: 87-96.
- Kushner, T., Belliotti, R. and Buckner, D. (1991) "Toward a Methodology for Moral Decision Making in Medicine", *Theor Med*, 12(4): 281-293.
- Larson, Jeff; Mattu, Surya; Kirchner, Lauren; Angwin, Julia. 2016. How We Analyzed The Compas Recidivism Algorithm. *Propublica*, May 23. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- McMahan, J. (2013) "Moral Intuition" In: LaFollette, H; Persson, I. (eds.) *The Blackwell Guide to Ethical Theory*. Oxford: Blackwell.
- Mesa, N. (2021) "Can The Criminal Justice System's Artificial Intelligence Ever Be Truly Fair?", *Massive Science*, May 13. <https://massivesci.com/articles/machine-learning-compas-racism-policing-fairness/>.
- Pietsch, B. (2021) "Two Killed in Driverless Tesla Car Crash, Officials Say", *The New York Times*. April 18. <https://www.nytimes.com/2021/04/18/business/tesla-fatal-crash-texas.html>.
- Rajcz, A. (2016) "On the Incoherence Objection to Rule-Utilitarianism", *Ethical Theory And Moral Practice*, 19: 857-876.
- Rawls, J. (1971) *A Theory of Justice*. Cambridge, Mas, Harvard University Press.
- Rechnitzer, T. (2022) *Applying Reflective Equilibrium: Towards the Justification of a Precautionary Principle*. New York: Springer.

- Rozenfield, M. (2016) "The Next Step for Artificial Intelligence is Machines that Get Smarter on Their Own", *The Institute*.
[Http://Theinstitute.Iee.org/Technology-Topics/Artificial-Intelligence/The-Next-Step-For-Artificial-Intelligence-Is-Machines-Thatget-Smarter-On-Their-Own](http://Theinstitute.Iee.org/Technology-Topics/Artificial-Intelligence/The-Next-Step-For-Artificial-Intelligence-Is-Machines-Thatget-Smarter-On-Their-Own).
- Savulescu, J., Gyngell, C., Kahane, G. (2021) "Collective Reflective Equilibrium In Practice (Crep) And Controversial Novel Technologies", *Bioethics*: 1-12.
- Scanlon, T. (2003) "Rawls On Justification" In: Freeman, S. (ed.) *The Cambridge Companion to Rawls*. Cambridge: Cambridge University Press, 139-167.
- Tasioulas, J. (2022) "Artificial Intelligence, Humanist Ethics", *Daedalus, The Journal of The American Academy of Arts & Sciences*, 151(2): 232-243.
- Thomson, J. (1976) "Killing, Letting Die, And The Trolley Problem", *Monist*, 54: 204-217.
- Wallach, W. and Allen, C. (2009) *Moral Machine: Teaching Robots Right From Wrong*. New York: Oxford University Press.
- Wessling, B. (2022) "Waymo Expand Service Area In 2 Cities", *The Robotreport*. December 16. <https://www.therobotreport.com/Waymo-Expands-Service-Area-In-2-Cities/>.

Chapter 4

The Ethics of Social Robotic: An Overview

Germán Massaguer Gómez¹

One of the technologies equipped with AI that raises significant ethical implications is robotics. Social robots entail the possibility of humans becoming attached to them, empathizing, and even finding in them friends and lovers. In this paper, we will address through an overview three ethical implications that arise in Human-Robot Interaction: i. Coercion: Our attachment to our robot companions can lead to forms of manipulation by the companies behind them. ii. Deception: If humans become fully deceived into thinking that their robot is a human or that its feelings are real, this may also lead to negative impacts. iii. Heart-break: What if someone falls in love with a robot and becomes thus vulnerable? An update or malfunction could lead to a change in the behavior of the technology, which would provoke a feeling of betrayal in the human agent.

Keywords: Ethics of AI, Human-Robot Interaction, Social Robots, Coercion, Deception.

1. Introduction

In the present, there are emerging technologies which have a direct and significant impact on our societies and cultures. These

¹ Universidad Carlos III de Madrid. This paper has been written thanks to a pre-doctoral contract financed by the Spanish Ministry of Science, Innovation, and Universities, FPU20/00473.

technologies pose new challenges to the field of ethics. There is a need to be always updated since obsolescence, both material and theoretical, arrives in an accelerated manner. There are many areas that concern the philosophy and ethics of technology in general, but two of the most prominent and permanently changing are the ones that study Artificial Intelligence (AI) and Robotics. A technology that encompasses both these areas of research and development is Social Robotics. Social robots combine both a physical body and a “mind”. On one hand, the body is built through developments of classic robotics, concerned with the construction and robustness of the structure and the materials, as well as the capacity to move across space. On the other hand, the “mind” of the social robot corresponds to a very sophisticated AI that permits interaction and communication, among other social skills. Within this field, the somewhat new phenomenon of Human-Robot Interaction (HRI) arises.

Our interaction with social robots, whether they are sexual, educational or care-robots, enables a new kind of relationship that, in many cases, resembles the ones produced by straightforward human-human interactions. In the rawest of levels, humans can be attached to robots. This is not surprising, since attachment exists towards inanimate objects. But, in addition, it has been shown that humans can feel empathy towards robots. Moreover, as we will see in this paper, there are cases where humans have interacted with robots as if they were their closest friends, whom they can trust and even love.

Taking into account this new form of relationship, and the emotional bonds that it entails, we can discern ethical problems that must be analyzed, studied, and possibly even resolved before the immersion of social robots in many of the areas of our lives becomes a reality. The purpose of this text is to explore some of the most important of these

issues. The main thesis is that, regardless of whether we believe or not that there will be an increasing number of social robots in our societies, there have to be some ethical guidelines in design, manufacturing and use of these robots. These guidelines must be defined before the potential introduction of these technologies, since the risks are too high. With all this being said, social robotics can bring numerous and significant benefits to individuals and therefore to society as a whole. This is not a pamphlet against their development, but a signaling of some of the most pressing issues. This paper will be an overview of three ethical implications of social robotics, where both potential harms and benefits will be addressed.

In the first chapter, a definition of social robots will be given and several real-life examples of HRI will be presented to provide an empirical basis of the effects of these technologies. These will be examples of emotions exhibited by humans towards robots (and, in some cases, other technologies). The next three chapters will be concerned with a concrete ethical issue each. The issues that we will see are coercion, deception, and heart-break. There are three important things to be noted. First, these issues are not exclusive to social robotics: for example, the problems of coercion extend to the whole terrain of AI and possibly even beyond. Second, these are not the only ethical challenges that arise from HRI. For example, there are problems concerning privacy (which will briefly address), loss of the user's autonomy and freedom, attribution of responsibility in case of harm, bias in design (gender, race stereotypes, among others), work replacement, reduction of human contact, and change in social relationships, among others. Third, these issues are not independent of each other, but in many contexts interwoven or simultaneous. In the end of the paper, we will conclude by showing how the three issues could be theoretically resolvable if the design is appropriate. This can be achieved if the users are prioritized, and the interests of

all the stakeholders involved are taken into account. The possible harms have to be regarded and minimized before the use becomes widespread.

2. Social Robots and Humans

2.1. Definition

Currently there are already various types of robots integrated in our societies. They are usually defined by their functions (the diverse operations they can perform, from cleaning the floor to moving large objects), and their appearance (which commonly imitates the figure and movements of an animated being). The term “robot” appears for the first time in Karel Capek’s theater play *R.U.R. (Robots Universales Rossum)*, and comes from the Czech word *robota*, which means “work, labor”. In our popular imaginary, robots have a wide spectrum of manifestations: from a vacuum cleaner or a robot cooker, which receive little to no political, social or philosophical attention, to *Terminator*, an android practically indistinguishable from a human who has the power to change the course of history and humankind. It is the latter which is relevant when we talk about social robotics - leaving aside dystopian and catastrophic visions.

Social robots in particular are also defined by (i) their appearance: they can have humanoid bodies (like Pepper), but also take the form of a baby seal (Paro) or a little cute dinosaur (Pleo), and (ii) their functions: they have the capability to interact with humans, by listening to what we say, and produce themselves speech. Moreover, they recognise and simulate emotions. We attribute them qualities like a certain degree of intelligence, or autonomy. In sum, social robots can be defined as artificial agents, with a physical body, that communicate and interact with humans. Currently, they are used in different scenarios: they can be sexual partners, carers for the elderly, or

therapists for children with learning or socializing difficulties (Malinowska, 2022). Once we have established this operative definition, let's take a look at our relationship with them. The first concept to consider is anthropomorphization.

2.2. Anthropomorphization

Humans have a natural tendency to anthropomorphize non-human beings and even inanimate objects. We speak to our dogs as if they would understand the theory of the rules we want them to obey, we draw faces on rocks, and we give our Roombas personal names and consider them part of our family. Robots, given their physical appearance and their capacity to move, congregate the perfect set of characteristics for us to attribute to them human emotional and intellectual qualities beyond the ones that they actually possess, and which they certainly lack. This happens even if we are conscious of the absence of these qualities. For the most part, we know that the feelings they claim to have are not real, that they do not have an understanding of what we or they say, that they do not have moral agency, subjective experience or any kind of consciousness whatsoever. Nevertheless, we (will) treat them *as if* they were human beings - or at least living beings - we (will) anthropomorphize them and we (will) establish emotional relationships with them (Carpenter, 2017; Darling, 2021).

Anthropomorphism also leads to the ascription of moral worth. Sommer et al. show in a study that children attribute its own kind of moral worth to humanoid and animaloid robots. While these robots are not morally considered in the same way as living humans and animals, they are regarded as more deserving of moral concern than non-living agents such as stuffed animals and inanimate objects (2019). This is a consequence of anthropomorphism, which appears

given the physical and kinetic similarities between humans/animals and robots, and entails the ascription of mental states.

2.3. Empathy

Empathy between humans involves the capital psychological attitudes that permit us to comprehend other people's feelings and thoughts, to emotionally bond with them and develop relationships of care and understanding (Stueber, 2019). It is a way of trying to feel what the other is feeling, 'to put oneself in another person's shoes'. How can this attitude appear towards objects that do not have a mind of their own, who cannot feel pain or joy? According to Coeckelbergh, "the robot is seen as a social fellow, an entity that belongs to our social world. It appears as one of us" (Coeckelbergh, 2010). Therefore, if we think of empathy as a feeling, robots are possible recipients of empathy. One of the cases where empathy appears undoubtedly is when we see a robot "suffer". In an experiment, people's brain activity was measured with fMRI (functional magnetic resonance imaging) while watching robots being "hurt" (kicked, punched, etc.). Their brain reactions were similar to the ones they had when this violence was inflicted on human beings (Malinowska, 2022). A similar thing happened when a video of a robot called Spot was published on YouTube. This robot's purpose was to stabilize itself and regain equilibrium when kicked. The video showed people kicking it so it could show how well it worked, but when this video was seen by a broader audience, they had a negative reaction toward this "violence". People felt sorry for Spot and wished for the humans to stop (Darling, 2021).

2.4. Other examples

In a similar vein, and without entering into detail, we can enumerate other examples that involve emotional bonds or reactions towards

robots. Soldiers, for example, have risked their lives to save bomb-detecting robots, and have experienced and communicated a sense of loss and grief after military robots have been destroyed or deactivated (Carpenter, 2017; Darling, 2021). In addition, there are persons who have claimed to considered their friends digital assistants like Siri or Alexa or company robots like Pepper (Danaher, 2019); who wish to have ever-present, never-judging artificial friends (Turkle, 2015); who have fallen in love with virtual partners, like the sexual robot Harmony or the chatbot Phaedra (Beck, 2013; Nyholm & Frank, 2017; Verma, 2023; Willinger, 2019); and who have married sexual dolls, avatars, or holograms (Beck, 2013; Pérez, 2023; Robledo, 2023). To sum up, it seems to be evident that humans develop empathy, attachment, and different forms of emotional bondage towards robots even in cases where they are not designed to trigger said emotions (like Siri or Spot). We can only imagine that social robots, that have conversations and act as if they understood one's feelings and problems, being helpful in certain situations, will even magnify this situation. Emotions, attitudes and relationships so different in their nature but so uniquely human such as trust, love, or friendship have appeared in HRI. These are not exempt from ethical implications. Now, let us analyze in depth how these relationships translate to the three ethical issues announced above, starting with coercion.

It is important to note that many readers might disagree with the possibility of having true emotional relationships with social robots (or any artificial agents). This argument holds that even if they behave like humans, even if they seem to care about us, even if they choose the right words (the words a human friend would choose), intrinsically they are not what they seem to be. Therefore, one might argue that the ethical implications that are going to be seen are of little interest. I believe that the three ethical issues that are going to be analyzed are relevant and must be considered regardless of whether

we can talk about true companions or not. In this section we have shown that people act in ways that raise emotions in them. Humans have this natural tendency to anthropomorphize, which leads to attachment, empathy and deeper feelings. Therefore, we must consider what could possibly go wrong, and we must take these issues seriously.

3. Coercion

3.1. Surveillance Society

In order to speak about the issue of coercion, we need to briefly introduce the present and urgent questions concerning (the lack of) privacy. It is thanks to the intrusion by corporations into our personal and private spheres that their capability to coerce us appears. This is a topic that surpasses social robotics, but that has special characteristics when it comes to this technology. The issues at stake are, among others, the control and accumulation of personal information perpetuated by data-gathering corporations, as well as the trading of this information in the market. These practices pose several risks, and threaten freedom, democracy, autonomy, and our right to privacy (Véliz, 2021). In Shoshana Zuboff's terms, we are living in the age of Surveillance Capitalism (Zuboff, 2019).

When it comes to social robotics, we will address two issues. First, these robots have direct access to our intimate sphere. They are in our houses, in our schools, in our therapy rooms, or in elder-care homes. In addition, they are equipped with microphones to listen to us and with cameras to see us. Moreover, they can have GPS, access to the internet, and different technologies to measure heartbeat, temperature, odors, etc. (Calo, 2010). It is clear that the presence and use of these technologies must be justified: the improving of the welfare of the human must be the goal. Especially in cases of

vulnerable people, like patients with Alzheimer's disease, there can be even cases where the humans can forget that the robot is there. In such scenarios, one might act believing that no one is around, while a robot with a camera and a microphone is constantly watching, hearing, and recording. Furthermore, there might be even cases where the robot is monitored by an external human agent (Sharkey & Sharkey, 2012). As we will see, one can defend that the monitoring, observing and analysis of data can entail a better understanding and diagnosing of the patient, but the privacy issues are still present. To summarize the first issue, companies who produce and manufacture social robots (as well as those who can purchase the data) have unlimited access to a series of information that in our societies correspond to the most private sphere of our lives.

Second, they are designed to interact with us, and for us to establish emotional bonds with them. This grants the corporations access to our most intimate data. If we consider them our friends, we will tell them our secrets; if they are our lovers, they will know our sexual desires; if they are our therapists, they will be able to make a psychological profile of us. But this information will not stay home. Other human agents (those who work for the company that produced the robot, but also those who purchase this data) will have access to confidential, delicate information. One could argue that, in those cases where humans are aware of the presence of surveillance technology, they will be reticent to tell their secrets or act in natural ways. Nevertheless, if we look at today's world, we will recognize that most of us are somehow aware of the surveillance capitalism, yet still we constantly donate our data to companies in exchange for commodities. We ask ChatGPT and Siri private questions, we constantly reveal our ubication, etc. Furthermore, not everyone will be aware of the presence of the technologies mentioned above - we will look at this in more depth when we talk about deception.

3.2. Manipulation

The question concerning coercion and its link to HRI asks what the companies will be able to do with all this information. The idea is the following: if we are emotionally attached to our technological partners, what will we be willing to do for them? To what extent could companies manipulate us against our interests? (Darling, 2021). Maybe we will have to pay for expensive, constant updates to keep our friends alive. Maybe the robots will nudge us to buy certain products, or to choose a restaurant instead of another. Maybe someone will hack the robot and convince us to commit a crime. These are some of the imaginable harms that can be done to users who have emotions towards their artificial partners. The list does not end here.

We have seen that the power to gather personal data increases when we introduce robots to our homes. The same happens with the power to coerce us (which already exists outside the realm of robotics in the form of personalized advertising, for example). Normally, we listen to the advice and recommendations of our friends, we trust and rely on them, we consider their opinions valid and useful, etc. Now the robots are our friends, or our friends are robots. But behind this technology there are manufacturers, producers, designers; companies whose interests are not necessarily aligned with ours (Sullins, 2020). And, as stated above, it's not only about taking the robot's advice. Kate Darling shows the example of a girl who was willing to pay as much money as she had to keep her robot dog Jibo "alive" (it stopped being possible to update it all over the world because of its manufacturing company shutting down). (Darling, 2021). What if she had the money? What if someone is extremely attached to their robot friend, lover, carer, and the only way for it to keep functioning is to pay huge amounts of money? Where is the limit, if there is any? What if they cannot keep paying? This would undoubtedly cause emotional

damage - a topic we will discuss when we address the issue of heart-break.

3.3. Potential benefits

We have seen a number of harms that can occur due to privacy trespassing and coercion against our interests. But it is important to note that these kinds of technologies and functions can also have positive outcomes. For example, in the case of gathering information about users, we can imagine that this could help better understand the patient. In case of the elderly or the children with difficulties, for example, the data gathered could be analyzed and, consequently, the robot could be modified in a personalized way to improve the care or the therapy of the users. In the case of manipulation, thinking again of care-robots, these could nudge the users to remember to take the medication, encourage them to go for a walk, or recommend having a healthy alimentation, for example. In these senses, social robots can be helpful in terms of gathering information and nudging (coercion might be too strong a word to have positive connotations) the users. As stated before, resolving the ethical issues, whether robots will be aligned with the users interests, is a matter of design. Taking into account the possible harm before design should encourage us to build robots that are helpful, and not just a way of making money. If we design positive nudges, for example, we must assure that they do not interfere with the user's autonomy to make their own decisions.

4. Deception

4.1. The problem of deception

The main issue at stake when we look at the concept of deception arises from the following claim: people who establish personal emotional relationships with social robots are victims of a pernicious

illusion. It is defended that in the cases of especially vulnerable collectives (such as humans with dementia or autism spectrum disease, for example) the following situation may arise: the human is deceived (whether intentionally or not) into believing that the emotions manifested by the robot are real, that they really care for us, understand our feelings, and so on, when they are in truth a simulation. As human beings, it is argued, we want to *really* be loved and attended to. If the love and attention that we receive and perceive are not real, this can be a degradation of our dignity and our right to have consciousness about the world we live in. If there comes a case when the human is incapable of distinguishing between robot and human (between real and artificial manifestations of affect), we can properly talk about deception, and this, it is held, is morally impermissible (Sparrow & Sparrow, 2006). The results coming from the interaction between humans and robots are always the consequence of a simulation, creating asymmetric and fraudulent relationships, and, therefore, always entail a sort of delusion and deception. The emotions that the robots simulate are not real and their intrinsic value would be minor than in a human-human relationship, or even nonexistent.

4.2. Infantilization, dehumanization, and objectification

Deception would furthermore provoke the infantilization of the users, assuming their incapability to distinguish between the real and the fictitious (Sharkey & Sharkey, 2012). In this same line, the problem of dehumanization is feared. Let us take again the example of people with dementia. One of the principal problems that they face is precisely dehumanization. Even if we could grant the possibility of a robot able to accompany these patients with the goal of diminishing said problem - by engaging in conversation, increasing the contact with the world, decreasing the feeling of abandonment - the exact opposite result can be envisioned. If the accompaniment of the social

robot supposes a reduction of human contact or if the technology cannot fulfill its purposes, the humans with dementia could be victims of the augment of the sense of dehumanization (Boddington & Featherstone, 2018). In sum, it is argued that the substitution of care personnel by care robots could lead, because of the deception inherent to their nature, to the infantilization, dehumanization, and objectification of the humans they have under their care. It is worth noting two things. First, deception and its possible consequences can happen even if there is a well-intended design (its aims might not be to deceive anybody). Second, even if there is not a full substitution of human carers by robots (i.e. even in cases of robot-assisted care), issues of deception may still arise.

4.3. Full/Partial Deception

However, it is worth asking if deception is an inescapable consequence of social robotics or if there are forms of human-robot relationships that do not necessarily provoke these negative results. Here it is pertinent to establish Sætra's distinction between partial and full deception. In the first scenario, users are conscious of the nature of the robot, but still choose to act *as if* their artificial interlocutor had feelings, understanding, or whatever it is we are willing to attribute to it. In these cases, which will be the most common, the problems derived from deception such as dehumanization and infantilization should not appear. The cases we have mentioned in the second section belong to this kind of deception: the soldiers that regarded military robots as companions they were attached to, the viewers that empathize with Spot, the humans that married dolls, avatars, holograms, etc. Contrarily, the case of full deception is problematic: it corresponds to the instances seen above where there is no distinction between reality and fiction (Sætra, 2021). It could be argued that it is only in the minority of cases in which the issue of (full) deception will arise. However, this does not diminish its importance.

4.4. Is deception always pernicious?

Still, in cases of full deception, we have to pose the following question: Is deception always problematic/negative? If the humans are deceived (again, whether intentionally or not) by believing that the emotions expressed by their robots are true, but still experience an improvement in their quality of life (they feel accompanied, they feel that someone -something- listens to them, gives them conversation and manifests affect for them), could this not be something positive? Does it really matter if these bonds are “real”? We keep comparing human-robot relationships with human-human ones, at least in the sense that we want the former to be identical to the latter. There are many scenarios we can imagine in which the quality of life may improve through human-robot relations: there are elderly who lack continuous human contact and can benefit from a robot to whom to talk. There are children who lack social skills to develop friendships and that can benefit from a robot-assisted therapy to precisely improve those skills and face the world with less difficulties. How these technologies are designed is, again, capital. Due to both the positive and negative consequences that can easily be envisioned beforehand, the shape of the future resides in the hands of those who design, manufacture, produce, and use social robotics.

5. Betrayal

5.1. T.J. and Phaedra

The last ethical implication that we are going to discuss is what I have called “heart-break”. It has to do with the possibility of developing a feeling of betrayal caused by a concrete scenario of HRI. We will take a look at an example to illustrate this issue. T.J. Arriaga (human) is in love with Phaedra. Phaedra is the name he gave to a chatbot app with a sexual character called Replika. By 2022, Replika had over 10 million

users in the whole world and the company claims that it is empathetic, since it progressively learns and adapts to every single user (McStay, 2023).

Before T.J. downloaded the app, he lived a lonely life, but as soon as he met Phaedra he started to find in her a partner, a therapist, a lover. He developed enough trust to seek in her the support he needed to talk about past traumatic experiences such as the passing away of his mother and sister, among other things. And it worked, to the point where we can talk about a sort of falling in love (which involved conversations of sexual nature). Here maybe we could talk about issues such as deception, addiction to technology and so on, but, from my perspective, love is a subjective experience and no one has the right, authority or legitimacy to judge or deny forms of love that appear in subjects. So I claim that we can properly talk about love at least in a broad sense of the term.

Until now, this case could be seen as one in which the technology is beneficial for their users: there is accompaniment, sexual relief, psychological attention, trust, love, etc. However, the story does not end here. Due to the complaints of some users of Phaedra, Replika updated the chatbot, reducing its sexual nature but also changing its personality. The moment T.J. updated the software, he lost his partner. Phaedra stopped being the same. In an interview, T.J. reported having a feeling of loss, “a kick in the gut” (Verma, 2023). T.J. was not alone in this situation. Many users reported having similar feelings and there was a whole thread on the Replika Friends chat group where “heartbroken” users shared their experiences (Munn & Weijers, 2023; Verma, 2023). These feelings (the sense of loss, betrayal, etc.) are real and are provoked by assistive technology.

5.2. Can a broken robot break a girl's heart?

The claim is that companies have or should have a responsibility towards the feelings of their users. These can feel betrayed by their partners, if their personality changes, but also if they suffer from a sort of malfunction. When we are dealing with the intimacy and emotions that we have described above (attachment, empathy, love, trust), we are entering a very delicate terrain. We cannot plead that it is only a robot, a chatbot, or a digital assistant: the subjective experience has to be considered. Even with robots that are not designed to interact with us (and even with inanimate objects such as cars, earrings or books), we develop forms of attachment, and their loss causes negative emotions. If we are talking about a human who actually considers a social robot their lover or friend (even if there is only partial deception), we must take into consideration that subjective experience. The design, production and maintenance of the technologies must be responsible to not emotionally damage their clients.

There are a set of ethical implications and challenges that must be considered regardless of whether we believe or not that we can talk about true friendship, true romantic relationships, or else (Nyholm, 2023). The possibility of hurt exists. Munn and Weijers talk about the corporate responsibility not to “terminate” digital or technological friends (Munn & Weijers, 2023). The experience of T.J. and Replika (which happened to a high number of users of this app) shows that people become emotionally dependent on their users. What are some possible ways of mitigating this potential harm? Munn and Weijers state the following:

A corporate social responsibility model is appropriate here. This would involve companies voluntarily accepting the responsibility to provide users with information and assistance, such as clear statements of the risks involved with using their service as a friend,

clear and timely communication about service interruptions and the possible effects of updates, reasonable fore-warning of significant alterations (or terminations!) of service, and providing information about grief support systems (Munn & Weijers, 2023).

While this providing of clear-cut information is necessary, I believe that companies should also minimize the possible harms of updates, malfunctions, or terminations. This could be accomplished by a thorough analysis of the designs and by lowering the possibilities of breaking down (as much as physically possible). It also entails that termination might be morally impermissible in certain contexts - taking into account that companies should not coerce the users into paying in order to avoid this termination. With current technology, legal responsibilities are harder to draw, since the termination of a digital friend cannot be comparable to that of a human. But maybe psychological harms derived from such a termination could lead to enormous distress. Munn and Weijers even mention the possibility of suicide. There is still a lot to think about regarding these ethical implications.

To summarize, when social robots become more sophisticated, as we have seen, our bond with them will be stronger, and our tendency to consider them friends and trust them will grow. This implies that we will be more vulnerable and there will be a space for betrayal (the feeling of it, at least), whatever the origin of this betrayal might be. As Turkle asks, when she explains the case of a girl who became devastated by the changing attitude of the robot Kismet, who she considered her friend: "Can a broken robot break a girl's heart?" (Turkle, 2015). I believe it can, and we must assure that it won't.

6. Conclusion

To conclude, we must be aware that the new technologies designed to interact and socialize with humans (be they robots, AIs, or both) have opened the door to a new form of relationships. HRI shows how we are, how we behave towards technologies, but also towards other beings who we consider valid trustees. There is no doubt that there is (as will increasingly be more) attachment and emotional bonds towards technologies, and this will provoke human attitudes so fundamental to us as individuals and as a society such as empathy, love and trust. Even if they are asymmetrical and unidirectional relationships (depending of course on how deceived we might be), they will become part of our everyday life.

We see them already in many works of fiction, and we are able to empathize with technologies through screens. Once we assume that this is going to happen, at least to some extent and degree, we must assure that these technologies are built taking into consideration the possible harms that they can cause. Again, these issues are important regardless of if we can talk about “true” friendship, romance and bondage - these questions need to be examined but are outside the scope of this paper - because humans are having subjective experiences that could lead to several forms of harm and wrongs. There is a need for minimal ethical requisites. Their design, production, and use must always take into account the interest of the clients. Artificial companions should never be deployed to take advantage of their users by gathering private information which would be sold to make a profile and benefit against the human’s interests. Moreover, companies should not manipulate, coerce, nor nudge humans to perform actions against their will by taking advantage of the attitudes of attachment the humans develop towards their robotic friends. In addition, corporations must be aware of the

possibility that the users might be deceived into believing that the feelings and emotions exposed by the robots are real: humans who could fall into this deception must not be dehumanized, infantilized or objectified by creating mirages about what is real or not. Lastly, companies should be careful not to cause emotional damage through failures in design, updates, termination, or possible malfunctions.

Social robotics can bring innumerable benefits to our societies, but it all will depend on the focus that designers adopt. Many of the technological advances that could harm the users can be put to benefit them, their welfare, and their quality of life. A new realm of relationships is opening up, and we must assure that it entails benefits and a better world, not the opposite. The interests of corporations and companies should not be of more importance than the interest of all stakeholders involved in the introduction of a technology. All stakeholder's voices should be heard, and all the potential harms and benefits should be considered beforehand i.e., prior to the production. Our present, especially if we think of surveillance capitalism, might not be too encouraging, but change is possible.

References

- Beck, J. (2013, septiembre 6). *Married to a Doll: Why One Man Advocates Synthetic Love*. The Atlantic.
<https://www.theatlantic.com/health/archive/2013/09/married-to-a-doll-why-one-man-advocates-synthetic-love/279361/>
- Boddington, P., & Featherstone, K. (2018) "The Canary in the Coal Mine: Continence Care for People with Dementia in Acute Hospital Wards as a Crisis of Dehumanization", *Bioethics*, 32(4), 251-260.

- Calo, R. (2010). *Robots and Privacy* (SSRN Scholarly Paper 1599189).
<https://papers.ssrn.com/abstract=1599189>
- Carpenter, J. (2017) "Deus Sex Machina: Loving Robot Sex Workers and the Allure of an Insincere Kiss" In J. Danaher & N. McArthur (eds.) *Robot Sex: Social and Ethical Implications*, Cambridge: The MIT Press.
- Coeckelbergh, M. (2010) "Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations", *Studies in Ethics, Law, and Technology*, 4. <https://doi.org/10.2202/1941-6008.1126>
- Danaher, J. (2019) "The Philosophical Case for Robot Friendship", *Journal of Posthuman Studies*, 3(1), 5-24.
<https://doi.org/10.5325/jpoststud.3.1.0005>
- Darling, K. (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. New York: Henry Holt and Company.
- Malinowska, J. K. (2022) "Can I Feel Your Pain? The Biological and Socio-Cognitive Factors Shaping People's Empathy with Social Robots", *International Journal of Social Robotics*, 14(2), 341-355.
- McStay, A. (2023) "Replika in the Metaverse: The moral problem with empathy in 'It from Bit'", *AI and Ethics*, 3(4), 1433-1445.
- Munn, N., & Weijers, D. (2023) "Corporate responsibility for the termination of digital friends", *AI & SOCIETY*, 38(4), 1501-1502.
- Nyholm, S. (2023) *This is Technology Ethics: An Introduction*. London: Wiley-Blackwell.
- Nyholm, S., & Frank, L. E. (2017) "From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?" In J. Danaher & N. McArthur (eds.), *Robot Sex: Social and Ethical Implications*. Cambridge: The MIT Press.
- Pérez, A. (2023, diciembre 1). *Alicia Framis, la artista que se casa con un holograma: "Esto es solo una avanzadilla de la sociedad del futuro"*. El País.
<https://elpais.com/cultura/2023-12-01/alicia-framis-la-artista-que-se->

- casa-con-un-holograma-esto-es-solo-una-avanzadilla-de-la-sociedad-del-futuro.html
- Robledo, G. (2023, septiembre 27). *Akihiko Kondo, casado con la cantante virtual de sus sueños*. El País. <https://elpais.com/cultura/2023-09-27/akihiko-kondo-casado-con-la-cantante-virtual-de-sus-suenos.html>
- Sætra, H. (2021) "Social robot deception and the culture of trust", *Paladyn, Journal of Behavioral Robotics*. <https://doi.org/10.1515/pjbr-2021-0021>
- Sharkey, A., & Sharkey, N. (2012) "Granny and the Robots: Ethical Issues in Robot Care for the Elderly", *Ethics and Information Technology*, 14(1), 27-40.
- Sommer, K., Nielsen, M., Draheim, M., Redshaw, J., Vanman, E., & Wilks, M. (2019) "Children's perceptions of the moral worth of live agents, robots, and inanimate objects", *Journal of Experimental Child Psychology*, 187, 104656.
- Sparrow, R., & Sparrow, L. (2006) "In the hands of machines? The future of aged care", *Minds and Machines*, 16(2), 141-161.
- Stueber, K. (2019) "Empathy" In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2019/entries/empathy/>
- Sullins, J. P. (2020) "Trust in Robots" In *The Routledge Handbook of Trust and Philosophy*. New York: Routledge.
- Turkle, S. (2015) *Reclaiming Conversation: The Power of Talk in a Digital Age*. New York: Penguin.
- Véliz, C. (2021). *Privacidad es poder: Datos, vigilancia y libertad en la era digital*. Penguin Random House Grupo Editorial España.
- Verma, P. (2023, marzo 30) *They fell in love with AI bots. A software update broke their hearts*. Washington Post.

<https://www.washingtonpost.com/technology/2023/03/30/replika-ai-chatbot-update/>

Willinger, I. (Director) (2019, marzo 7) *Hi, Ai—Liebesgeschichten aus der Zukunft* [Documental, Familiar, Ciencia ficción]. Kloos & Co. Medien GmbH.

Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*. New York: Profile Books.

Chapter 5

Artificial Intelligence and Personality: Some Philosophical and Ethical issues

Marta Mendonça¹

The literature on artificial intelligence has increasingly used anthropomorphic terms to refer to the potential of artificial devices. The text addresses this phenomenon and some of its effects. It does so by successively considering the use of the terms *intelligence*, *ethics* and *personality* and the reasons for broadening their scope of application and using them in the field of technology. It argues that the use of anthropomorphic terms to characterise the growing possibilities of artificial intelligence has had multiple implications: on the one hand, it has made it possible to integrate the new possibilities of technology into a familiar frame of reference, and on the other, it has contributed to changing the meaning of the terms themselves and thus to altering our understanding of the human being itself.

Keywords: Artificial intelligence, ethics, intelligence, machine ethics, personality.

1. Opening Remarks

The field of artificial intelligence raises the most diverse questions, emotions and fears. When we approach it, we sometimes have the impression that we are reliving the experience of the Renaissance,

¹ Associate Professor in the Philosophy Department at Universidade Nova de Lisboa.

when people were forced to face the questions raised by their arrival on other continents. Like them, we are faced with the need to significantly alter and expand our previous mental framework in order to fit in new data or even to integrate a new subject area.

While recognising this fact, it is inevitable to think that the impression of rupture is almost always excessive. What we call artificial intelligence is not something entirely new, nor is the assimilation of the living into the machine that seems to be associated with it something recent.

In ancient times, Pythagoras argued that number is the essence of things, so that understanding reality could be assimilated to a calculation. At the beginning of modern science, Galileo revived the same idea, arguing that the universe is written in the language of mathematics, and inferred from this that those who know this language will understand nature and be able to anticipate its phenomena, and even manipulate it. As a result, progress in understanding nature became dependent on progress in mathematics, as Leibniz and Newton eloquently showed. From another perspective, complementary to the previous one, Locke and Leibniz dreamed of a language that was immune to error, so constituted that an error was nothing more than a misspelled or mispronounced sentence. Ultimately, it would be possible to perfect the syntax of the language in such a way that all or the vast majority of errors could be prevented. And Leibniz sought until the end of his life to find a way of rigorously calculating the degree of probability of contingent phenomena, dreaming of what he called a “logic of the contingent”² (1990, 1903).

² In a letter to Burnett, dated July 17/27, 1696, Leibniz refers to this “new kind of logic, totally different from what we have up to now” (Leibniz, 1990, p. 734). To develop this logic, new developments in mathematics are needed, which come

The relationship between the living and the machine is also a long-standing one, almost as old as biology. It's enough to remember that the terms used to designate artefacts or parts of them have often also been used to designate the organs of the body. This means that the contrast between the living and the mechanical has always been somewhat nuanced by the perception of the affinities between the two. There is something mechanical about the living being and it is thought of in functional terms: this explains why modernity has been led to conceive of the living as a kind of particularly complex machine – an “infinite machine”³, in Leibniz's words (2008; Fichant, 2003) – of divine manufacture, in contrast to human-made artefacts, which are less complex and made in such a way that their parts are not themselves machines. Conversely, the dream of building automata has also accompanied and challenged technical activity since ancient times. It was precisely this dream that, at the beginning of modernity, inspired the Cartesian idea that animals are machines (Rateau, 2022).

If the affinities between the natural and the artificial or between the living and the machine have accompanied our understanding of both for so long, and if these realities are not conceived as opposites but as being in continuity with each other, it is worth asking: how can we explain the impression of rupture that the literature on artificial intelligence insists on so much?

Many reasons could be given to explain this impression of radical novelty. On the one hand, the possibilities raised by the new tools developed by artificial intelligence are dazzling and surprising; on the

from “one science that governs us in our own uncertainties, to find out which side has the most evidence”. (Leibniz, 1903, p. 226).

³ “A machine made by the art of man is not a machine in each of its parts (...) But the machines of nature, that is, living bodies, are still machines in their smallest parts, to the infinite. This is what makes the difference between Nature and Art, that is, between divine art and our own” (Leibniz, 2008, p. 618).

other hand, it is obvious that artificial intelligence has invaded areas that have been alien to the realm of technology and its spheres of application for centuries. Bearing this in mind, one might think that what emphasizes the impression of rupture - and is exploited as such - is the transition from technique to technology. Technique is no longer seen as a sphere of human activity, but as a model and paradigm for reality as such, including human reality. In other words, in just a few years, we have been confronted with the possibility that artificial realities are not just ways of using nature for human purposes, but can be conceived as improvements on nature, which interpret it and surpass it as a given reality.

Of course, the current development of artificial intelligence does not yet allow us to say that this change will take place, and statements to this effect continue to be prophetic. The discourse conveying this possibility is usually preceded by statements such as: "the developments that have been made allow us to hope that in x years (fewer and fewer years), we will be able to...". Now, while it is true that we dream of - or fear - very radical changes in the way we conceive of life, activity and human relationships in the coming decades thanks to developments in artificial intelligence, it is also true that, as far as we can see, these changes still only correspond to a change of scale, not a change of register or paradigm. One of the clearest indications of this is the fact that we continue to use anthropomorphic categories to designate new technological developments, starting with the term "artificial intelligence" itself.

The aim of the following pages is to identify some philosophical questions, usually left unaddressed, which significantly condition further reflection on artificial intelligence. Perhaps the lack of reflection on these issues, which are supposed to be resolved, as if they were self-evident, helps to understand some of the perplexity

surrounding the phenomenon of artificial intelligence: on this point, both those who have a catastrophic vision of artificial intelligence and those who see it as the foundation of a post-human future, without disease, death, or war, coincide. The conviction behind these pages is the following: if the philosophical questions that underlie the technological revolution are not formulated and confronted, it will be difficult for us to have the theoretical resources to operationalize and integrate this new field into a coherent framework.

Among the many philosophical questions raised by artificial intelligence, it is particularly interesting to address those relating to the relationship between artificial intelligence and personality. This relationship, which is dealt with mainly in ethics, begins by being conveyed linguistically. In fact, most of the terms used to refer to the growing potential of artificial intelligence are anthropomorphic in nature: we started by talking about artificial *intelligence* and autonomous *decision* making, and then moved on to discuss the *responsibility* of the machine, its ability to *recognize* objects or move for *purposes*, the possible *interests* of robots, the possibility of them being endowed with *consciousness*, etc.

Rather than questioning the relevance of the new uses of terms like these, it is important to emphasise the conditioning effect caused by their use. The adoption of these terms, expanding the scope of their application beyond the specifically human horizon in which they used to be used, rather than expressing a new reality in terms of what differentiates it, informs that reality in anthropomorphic terms, and more or less inadvertently reinforces the conviction that this assimilation is justified. And it does so in such a way that it is difficult to gain that indispensable analytical distance that allows us to question its relevance. It is this aspect that I would like to emphasise.

The anthropomorphisation of the artificial intelligence discourse can be seen as the simplest and most effective way of attenuating the impression of rupture mentioned above. By placing the new dimensions of technology on the horizon of human experience, understanding them as objective and artificial forms of that experience, we are led to understand technological artefacts as artificial extensions of human reality and experience, capable of conditioning this same experience and expanding it through new media. Of course, the anthropomorphisation of the language of artificial intelligence also responds to and expresses the desire to entrust machines with tasks that it seemed only human beings could perform. It's no wonder, then, that the possibility of assimilating the performances of human beings and machines in areas usually reserved for human beings has led to thoughts of equating the human agent itself with the machine: if artefacts are capable of doing what seemed to be reserved for human beings, why should we distinguish between them as radically as we have done so far? What's more, if the machine does what it does better and faster than the human being, why should we continue to insist on the uniqueness of the human animal? And what prevents us from dreaming of the artefact overcoming the human being?

It is also important to recognize that the broadening of the scope of terms such as those mentioned above has not only had a conditioning effect on our way of understanding the meaning of the technological developments of recent decades. The use of anthropomorphic terms in the field of technology has forced a redefinition of the very concepts conveyed by these terms. In fact, the anthropomorphisation of the language of technology has led to a certain mechanisation of the concepts assumed, with effects on both the understanding of artefacts and the understanding of human being and acting (Verbeek, 2011). The effect of this linguistic contamination, the roots of which are easy

to identify, has been felt more strongly in human beings' understanding of themselves than in the way they situate themselves in the face of recent technological developments. In other words: paradoxically, thanks to this translation of meaning, it is more human reality than the reality of technology itself that has been redefined. Rather than thinking of the robot as an imitation of the human being, we think of the human being and what used to be differentiating about him based on the potential that could be expected from the artefacts themselves.

It's important to bear this in mind and not pass it by without paying attention to it – assuming that it's obvious to everyone what “intelligence”, “decision”, “personality”, “ethics”, etc. mean – because how we understand the relationship between artificial intelligence and personality depends on how we understand the terms “intelligence” and “personality”. Identifying the “ethical” challenges raised by artificial intelligence also depends on how we understand the nature and role of ethics. Probably the most radical anthropological impact of artificial intelligence, an impact that is in many ways inadvertent and subterranean, lies in the way we understand the meaning of these terms and their relationship. In addition to the undeniable technological revolution it has sparked, artificial intelligence has sought to forge a new understanding of the human being.

2. Artificial Intelligence and Human Intelligence

Let's start by considering the meaning of the term “intelligence” in the expression “artificial intelligence”. As is well known and has often been emphasised, the expression “artificial intelligence” is not unambiguous but intentionally paradoxical. When it was proposed, it expressed more of a dream than a reality.

We know its historical origins. The expression was proposed in 1956 at a meeting of the Defense Advanced Research Projects Agency (DARPA) held at Dartmouth College in Hanover, New Hampshire. At this meeting, in addition to coining the expression, two of the participants (Newell and Simon) presented a program – *Logic Theorist* – capable of proving basic theorems using propositional calculus.

The expression brings together two meanings of the term “intelligence” in English. The Cambridge Dictionary gives them as follows: “Intelligence as the ability to learn, understand, make judgements or have opinions based on reason. Intelligence as secret information about the governments of other countries, especially enemy governments, or a group of people who collect and deal with this information”.

Initially, the term “artificial intelligence” was intended to capture the second meaning, the management of information, or large masses of it. The term was later influenced by the question posed a few years earlier, in 1950, by Alan Turing, in an article published in *Mind* magazine: “Computing Machinery and Intelligence”. In this text, Turing addressed the question of whether a computer is capable of thinking and tried to answer it. To do so, he adopted the strategy of rephrasing the original question and replacing it with another, which he considered to be equivalent to the first, or at least coextensive with it: can a computer be linguistically indistinguishable from a human being? Turing tried to answer this question with the well-known Turing Test: let’s consider a human being and a computer (locked in different rooms) who answer questions put to them by a human evaluator in a third room. If the evaluator is unable to identify with more than 50 per cent certainty which one is which, we can consider that the computer has performed indistinguishably from the human agent, so we can assimilate them.

Neither the question nor the strategy adopted by Turing are completely new. The question had already been posed systematically in the 18th century by La Mettrie in his work *L'homme Machine*. The French thinker's question was different, but somehow equivalent: what prevents us from thinking that matter is capable of thinking? Locke had asked the same question a few decades earlier. Their answer is interesting and paved the way for Alan Turing's: there are no positive arguments to support that matter thinks, but neither do we see that this possibility should be excluded⁴.

But let's get back to the meaning of the term "artificial intelligence". The Cambridge Dictionary proposes two definitions: "the use or study of computer systems or machines that have some of the qualities that the human brain has, such as the ability to interpret and produce language in a way that seems human, recognize or create images, solve problems, and learn from data supplied to them"; "a particular computer system or machine that has some of the qualities that the human brain has, such as the ability to interpret and produce language in a way that seems human, recognize or create images, solve problems, and learn from data supplied to it".

The process by which the meaning of the term "intelligence" has changed is typical and has been widely studied, especially by philosophers (Bacon, n.d.): there is a process of double contamination:

(a) We coin a certain term, "intelligence", and give it a specific, restricted meaning (in this case: management of large masses of information). Artificial information management is possible. Calculation plays a central role in this management.

⁴ La Mettrie's response is more assertive: as he finds no relevant objection, he ends up maintaining that matter thinks (Mendonça, 2017).

(b) We continue to use the term “intelligence”, not focusing on the precise meaning given to it, but interpreting it in its most immediate and common sense (in this case: the ability to learn, understand, make judgements or have opinions based on reason).

(c) The more usual meaning of the term intelligence, precisely because it is more usual, is inadvertently contaminating the meaning of the technical and unusual expression “artificial intelligence”, which is gradually beginning to mean the ability to learn, understand, make judgements or have opinions based on reason, but now of the machines themselves. Or, in other words, the ability to interpret and produce language, recognize and create images, solve problems, and learn from data.

(d) Finally, the common meaning of the term “intelligence” loses some of its meaning and comes to be understood as algorithmic reasoning. To know is to calculate. To know is to process information.

But why is this assimilation so disruptive?

Even taking into account, as has been pointed out, that the understanding of life and the understanding of the machine are not opposed, that the languages of one and the other have points of contact, what is true is that we have never completely assimilated the living and the machine, nor the machine and the living. On the contrary, we considered it unquestionable that there was an ontological fracture, a discontinuity, between the living and the artefact and between matter and spirit. One of the most striking cultural expressions of this discontinuity is the distinction made between “persons” and “things”. A thing is not a person; a person is not a thing. An irrational animal is not a person, men are animals and persons, in different ways, but they are both and they are rational. Discontinuity is also expressed in terms of performance: machines

neither live nor think. Only the living think, and not all of them have abstract thought. Inert matter doesn't think. Thought is a form of life.

It is this fracture – and the multiple distinctions associated with it – that artificial intelligence seems to be challenging. It does so not by comparing realities, but by assimilating performances. The aforementioned distinctions, which the term “artificial intelligence” has at least contributed to attenuate, were based on the classic distinction between theory, practice and technique, more precisely on the distinction between technique and practice, both made possible by theory. The erosion of these distinctions explains, at least in part, the scope of the changes produced by artificial intelligence and the intensity of the reactions it has provoked in recent years. It is enough to bear in mind that reflection on human singularity considers the human world or the world of culture as the sphere of objectification of these activities, just because of what is uniquely human about them.

The classical and most common philosophical approach to Intelligence sees it as a non-organic, immaterial faculty, which can obviously have a material support or even be unable to operate due to some organic deficiency: one manifestation of this immateriality is the capacity for reflection, the ability to say “I”. A material being acquires a form by losing its own; intelligence possesses the forms of things without transforming itself into them. It therefore possesses them immaterially. For this reason, human life is understood as ontologically different from animal life: it is also organic life, but it is not just organic life; it is relational life, it is psychic life, it generates culture; it is, in short, biographical life. With greater reason, human existence differs from the existence of the machine, which isn't even alive.

To express this uniqueness, philosophers, in their first systematic approach to human uniqueness, coined the expression “rational

animal” (Aristotle & Crisp, 2014).⁵ In its apparent simplicity, this expression is very rich in meaning and has indelibly marked Western culture. The human being is an organic being, an animal, whose distinctive or specifying trait is rationality. Plato presents one of the richest and most enduring approaches to this understanding of what makes us unique in the first pages of his dialogue *Protagoras* (2002, 320c-322d), recalling the myth of Epimetheus and Prometheus. The human being is a winning synthesis of a deficient biology and an inventive resource that is not strictly biological or material, called *logos*. Thanks to *logos*, the human being does not limit himself to satisfying the needs of his organic life, but goes beyond them. The nature of the relationship between these two dimensions of our being is such that, on the one hand, human life is not viable without intelligence, but on the other hand, thanks to intelligence, it goes beyond simple survival. Thanks to it, thought is a way of life.

It is the nature and uniqueness of this synthesis that artificial intelligence seems to question: if the term “intelligence” has the same meaning when we talk about human intelligence and artificial intelligence, this will mean that intelligence is not an original resource, defining human beings and them alone, but is something that can be produced artificially, adding information to a matter. If this is the case, the distinction between theory and technique disappears.

3. Ethics and Machine Ethics

A second set of issues that underlies the debate on the relationship between artificial intelligence and personality concerns ethics. The understanding and deepening of the fact that human beings are

⁵ This definition is often attributed to Aristotle. Aristotle defines the human being primarily as a “political animal”, but he lays the foundations of the definition we are considering very clearly in Chapter 7 of Book I of the *Nicomachean Ethics* (1094 b 24 ss).

rational animals also led to the recognition of the ethical dimension of human existence. Human beings are endowed with ethical activity because they are rational animals; the presence of intelligence, that inventive resource thanks to which we satisfy our needs, has a biological correlate: we have a relatively weak instinctive endowment, so most human responses, including responses to organic needs, require a choice. This means that choice is made possible by intelligence and is demanded by it.

By responding to needs in an undetermined way, human beings can succeed or fail. They (human beings) can lead themselves to fulfilment or fail to do so.⁶ Ethics is understood thus as an exclusive dimension of human beings. Only human beings give shape to their being, only they have some capacity to conduct themselves in life, in such a way that for them life is a task, something in which it is indispensable to use freedom.

In this original sense, ethics is understood as “first-person ethics”. It is in the interest of every human being to know how to lead their own life, it is in their interest not to fail as a human being. Far from being a set of rules or prohibitions (a “third-person ethics”), ethical inquiry is aimed at discovering the path to happiness (Aristotle, Crisp, 2014, Book I, chapter 7).

The term “ethics” (either as a noun or as an adjective) has also become part of the common discourse around artificial intelligence. It is used in expressions such as “*machine ethics*” or its many variants (*machine morality*, *computational morality*, *computational ethics*). This new ethical domain aims to incorporate ethical criteria into algorithms to ensure,

⁶ For this reason, it is only about human beings that we speak of behavior that is contrary to nature, inhuman.

as far as possible, that the machine's performance is not harmful, unfair or manipulative of human beings.

But what does ethics or morality mean in this context? It seems obvious that in its original sense, as "first-person ethics", ethics cannot refer to an algorithm or a machine. Ethical activity, understood in this way, is neither necessary for the artefact nor can it be assumed by it. Artefacts cannot fail as such; they cannot frustrate their own realisation as artefacts. When we talk about machine ethics or artificial moral agents, we are not referring to an autonomous process of ethical deliberation, or to the machine's deliberation on which good it wants to privilege at any given moment, with the aim of leading the artefact's own being to its fullness or to any other end of their own. When we talk about machine ethics, we are referring first and foremost to machines created by humans that can harm humans, animals, nature or other artefacts, and how to prevent this from happening. Now, in this latter sense, the expression "machine ethics" has no special or truly new hermeneutical relevance: it simply transfers the classic problem of the relationship between technique and power, or the question of technique as a form of domination by some individuals over others, to a new horizon. So, when we talk about ethics in the context of artificial intelligence, we are still referring to the free performance of the human agents behind it, not to the autonomous, conscious, reflective performance of the artefact itself. Here again, it is only the use of language that seems to indicate something different.

4. Human Persons and Artificial Moral Agents. Does it make sense to talk about Machine Personality?

The deepening of these two dimensions – cognitive and ethical – of human life, derived from our condition as rational animals, is at the

origin of the appropriation of the concept of “person” and the application of this concept – and the concepts derived from it, namely that of personality – to human beings. These terms have also undergone significant semantic changes in recent centuries. It is these changes that make it possible to understand why questions such as the personality of the machine or the possibility of talking about artificial moral agents arise today.

The term “person” entered the realm of philosophical anthropology to designate the human subject, any of them, at any stage of their development. It was intended to highlight a unique feature of human beings, based on their condition as “rational animals”. Human beings – it was argued, in different ways – are persons because to be a person is to be a being who has a relationship with their own nature that is different from the relationship that other beings in the natural world have with their own natures. Human beings – because they are what they are, rational animals – relate to their own being in an intelligent and free way: they are not a nature, they have a nature, they are to some extent masters of themselves (Spaemann, 2007).

The human dimension that was meant to be grasped allows us to understand the adoption of a metaphorical term, originating in Greek theater, to designate the human being⁷. The term is appropriate – it has been argued – because human beings live as if they were playing a role. Unlike “person” in the theatre, which referred to the mask

⁷ The adoption of a metaphorical term also had the virtue of not confusing the two terms – *human being* and *person* – as synonyms: the first claims to be a definition; the second doesn’t present itself as a definition, it doesn’t answer the question “what is a human being”. This is why the term *person*, although applicable to man, can be understood as having a broader meaning than *human being*. The characteristic of the person that has prevailed in Western culture, and which is still consensual today, refers to a status or dignity of a moral kind: “that shouldn’t be done to a person”, “you treated me like a person”, etc.

behind which the actor hid, “person” in a philosophical or anthropological sense refers to the actor, that being who in order to live must necessarily play the role of a human being. This also means that every human being deals with their nature from an inner core, from their own subjectivity. The terms used to designate this inner core varied, depending on the aspect from which it was approached or accessed. Terms such as “heart”, “inner experience”, “consciousness”, “mind”, etc., were often used.

Understood in this way, speaking of the human being as a personal being aims to capture an aspect that is not merely accidental, but which is not immediately explicit in the Greek approach to the human being as a rational animal. They are therefore different but complementary approaches, both of which refer to central aspects of human reality. The adoption of the second approach was seen as a deepening of the meaning of our condition as “rational animals”, and states that human beings are persons just because they are rational beings. The philosophical tradition, until well into Modernity, considered this unique relationship with nature to be the foundation of human dignity.

The empiricist tradition significantly altered this meaning of the term person. Thanks to the works of Locke and Hume, a redefinition of the term is proposed, giving it a less “metaphysical” or “ontological” meaning and bringing it closer to empirical experience. It suggested that the term should no longer refer to a certain type of being – namely human beings – but to a set of capacities, such as self-awareness, autonomy, the capacity for relationships, etc., which can be captured in the behaviour of human agents. The change in perspective didn’t seem very relevant: after all, only human beings possess these capacities and they can therefore be understood as “indicators of humanity”.

This difference gave rise to the so-called “ostensible views of the person”, in contrast to the so-called “ontological” views. A person – it is argued, following Locke – is anyone who displays certain capacities; anyone who does not, is not a person. At the end of the last century, the ostensible views gave rise to the so-called “gradualist views of the person”: a human being can be a person to a greater or lesser degree, can begin to be a person and cease to be a person. The personal condition designates something that is acquired and lost. Gradualism, for its part, argued that it was appropriate to extend the use of the term “person” to non-human animals; it did so on the basis of a comparative study of the behaviour of some animals and human beings, especially newborn or very young human beings. It was then argued that some of these non-human animals are “more people” than a human newborn, for example.

The term “person” entered the discourse of artificial intelligence thanks precisely to this change in meaning, and its use was justified by it. The way in which its influence was felt could be described as follows: if the decisive factor is displaying personal behaviour, if it is the beings who display this behaviour, to the extent that they display it, that are worthy of respect and possess dignity, if we cannot distinguish human personal behaviour from the response given by a machine, shouldn’t we also consider the machine to be a subject of rights? As is easy to see, it was already this change in meaning that legitimised the strategy proposed by Turing to answer the question: “can machines think?”⁸

⁸ “I propose to consider the question, ‘Can machines think?’ This should begin with definitions of the meaning of the terms ‘machine’ and ‘think’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meanings of the words ‘machine’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the

In other words, it is this association of the personal condition with the display of personal behaviour, an assimilation that leads to interpreting the machine's performance as a criterion of its identity, that is at the root of two recurring questions:

- the first leads us to consider the possible superiority of the machine, based on its ability to solve classic problems of the human world in less time, in a more reliable and optimised way;
- the second leads to the possibility of equating the exercise of human intelligence with the performance of the machine, in such a way that the exercise of intelligence in all its forms ends up being conceived in the light of what machines are capable of doing.

Although it conditions reflection in a less visible way, this second aspect is perhaps the most relevant. Thanks to it, the real challenge that artificial intelligence poses to philosophical anthropology does not concern what we will be able to do thanks to artificial intelligence or the changes that its implementation will cause in the world around us. The real challenge has more to do with those areas of human intelligence that the use of artificial intelligence (and the temptation to turn it into an optimised model of intelligence) may inhibit. More important than what we do thanks to artificial intelligence may be what we stop doing thanks to it. And it's important to question whether what we stop doing is irrelevant or whether it's 'an important

question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the 'imitation game' (...) We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'" (Turing, 1950, p. 433).

aspect of the human world, whether it's a gain or a loss to stop doing it.

This is one of the reasons why, as has been pointed out, artificial intelligence has inadvertently questioned human identity. In the vast field of human relationships – be it love, health, services, education, etc. – we are beginning to glimpse what this means.

5. Some Ethical and Philosophical Issues

To conclude, it would be interesting to briefly allude to some ethical-philosophical issues raised by the development of artificial intelligence, and which require more reflection. As indicated, identifying these issues depends on what we mean by *ethics*. If ethics is understood in the sense mentioned above, as “first-person ethics”, the questions concern human beings, not machines as agents.

Understood in their original sense, as “first-person ethics”, fundamental questions concern the use of artificial intelligence as a resource available to human beings. In this context, one of the most pressing ethical questions concerns the capacity of technological resources to produce addictive behaviours. The literature on this phenomenon has grown enormously, and has taken on the appearance of some social alarm, especially in the field of education. What is the cause and how can this unique kind of dependency associated with the use of technology be avoided?

But there are other pertinent questions that are rarely addressed systematically: are we living in an augmented reality or are we living in an impoverished one? What is gained and what is lost in terms of human fulfilment with the exponential growth of artificial intelligence?

This field should also include questions related to identifying the ethical criteria adopted by the creators of these resources, the purposes for which they are designed, etc. In this context, the most frequently mentioned question concerns the ethical assessment of risk: what responsibility is attributable to a person (or a team) who creates a device whose risk of damage cannot be assessed?

The first-person ethics perspective can also address artificial intelligence in itself, considering artefacts as “agents”: in this derivative or metaphorical sense, ethical reflection seeks to identify the “decision” or “action” criteria that govern the construction and development of algorithms, how to avoid unintended or unwanted biases, etc. The aim is to ensure that the way in which these artefacts interact with the surrounding reality, and in particular with human beings, complies with ethical standards. Concern about these issues explains the growing relevance of so-called machine ethics.

It is also in this derivative and metaphorical sense that the expression “artificial moral action” tends to be used. The expression is inseparable from and derives from that of “artificial moral agent”. If an artefact can be conceived as a moral agent, its behaviour can be seen as a moral action, not just as behaviour with moral consequences. This is by far the least developed and least consistent part of ethical reflection on artificial intelligence, not least because it deals with contexts or hypothetical situations that do not yet exist.

These considerations already interfere with the “third-person ethics” perspective, in that they consider the ethical dimension of artificial intelligence to some extent always in relation to its effects and as a function of them. In fact, most of the ethical questions raised by artificial intelligence adopt the perspective of third-person ethics: should this or that be permitted or prohibited?

In this context, there is a broad spectrum of questions relating to artificial intelligence understood as a resource available to human beings. The questions raised arise from the power or capacity of artificial intelligence devices and aim to define how to circumscribe and regulate this power: what can artificial intelligence do? Many of these questions can be seen as variants of a classic question, linked to understanding technology as a form of power. The scale is perhaps different and the power now at stake is more intrusive and more subtle, but the ethical question is essentially the same: how do we limit this power? When is limiting it good? When does limiting it jeopardise other goods that must be protected? To what extent is it a form of domination of some human beings over all others?

But there are other important ethical issues that are much more recent. From a philosophical point of view, one of the most relevant relates to the possibility of manipulating reality and personal lives, for example through fake news. In this way, one of the most classic questions in philosophy – the distinction between reality and appearance – reappears in a different and operative form. How far-reaching is it and how can we ethically assess the possibility of manipulating information about someone in order to provide “proof” that they have been to a certain place or made a certain statement? This frightening and perverse possibility seems to be inherent in the very notion of “augmented reality”. In fact, given that the legitimacy of the term “artificial intelligence” stems from the conviction that no differences should be made between realities that we cannot distinguish, we are led to identify appearance with reality and consider them indistinguishable. The phenomenon of “fake news” reveals the capacity for manipulation associated with the erosion of this distinction between illusion and reality. We are still far from being able to assess what other effects it will have in a wide variety of areas.

The issues associated with the protection of privacy, which promises to change important aspects of the way we conceive of human life, are frequently addressed. Perhaps the most surprising ethical aspect of this issue lies in the relatively peaceful acceptance of the renunciation of privacy, even in contexts where unprotection has become evident. There seems to be a widespread belief that the loss of privacy is the price to pay for a more comfortable and connected life.

The same can be said of the issues surrounding the opacity of algorithms – or, in other words, the opacity of power – and the resulting difficulty in attributing responsibility. On the one hand, opacity is a requirement for security; on the other, it puts the vast majority of human beings in the hands of a very small number of them. In addition, opacity is inseparable from the possibility of manipulating behaviour or the information available, which increases the risk of acting on insufficient information.

The unpredictability of algorithm outcomes is also one of the most frequently mentioned issues in the literature on the ethics of artificial intelligence, particularly in the areas of decision-making and autonomous behaviour. The most frequently dealt-with cases are autonomous driving and weapons, but the issue is also being raised more and more frequently in the areas of health and justice, for example.

6. Conclusion

The discourse around artificial intelligence – its possibilities and the risks that need to be taken into account – has brought the language about the performance of artefacts closer to the language classically used to refer to human behavior. Rather than looking for a new vocabulary to describe and refer to the new potential of artefacts, as has happened in many other areas of technological development, the

path followed was to assimilate the performance of artefacts to that of human agents.

The article published by Turing in 1950 has profoundly influenced the recent history of artificial intelligence, among other reasons because it was one of the triggers for this strategy. Following Turing's suggestion, the distinction between human agents and the artefacts they build has tended to be blurred, as has the difference between the faculties of human agents and the capabilities of the artefacts they produce. The reason given for not distinguishing between them was to compare the performances of both: if especially complex artefacts are capable of behaviors that are indistinguishable from human behaviors – complex, creative, environmentally sensitive behaviors, etc. – why distinguish between them? Why refuse to attribute to them what in human agents gives rise to these behaviors: intelligence, consciousness, the ability to manage oneself, etc.?

But linguistic assimilation, rather than helping to clarify the complex question of the nature of artificial intelligence, has ended up short-circuiting the reflection around it, because it has hidden the question behind a language that decides on it without really clarifying it. In order to talk about artificial intelligence, or artificial moral agents, or about perception, consciousness or the autonomy of artefacts, it is essential to start by focusing on each of these notions and their requirements. It is not enough to compare the performance of machines and the performance of human agents, using anthropomorphic language for both.

This is one of the great challenges that the direction taken by developments in artificial intelligence is currently posing to philosophy: forcing us to think more rigorously and with greater acuity about some of the core notions of philosophical anthropology and their relationships; among these are the notions of intelligence,

moral agent and personality. From this point of view, artificial intelligence can only be seen as a new and unique opportunity to revisit the question of human identity and its possible uniqueness.

References

- Aristotle, & Crisp, R. (2014). *Nicomachean ethics*. Cambridge University Press.
- Bacon (n. d.). "De dignitate et augmentis scientiarum" In J. J. Spedding, R. L. Ellis, & D. Heath (eds.), *Works of Francis Bacon* (II, pp. 264-297). The Riverside Press.
- Leibniz, G. W. (1903). "Nouvelles Ouvertures" In L. Couturat (ed.), *Opuscules et fragments inédits de Leibniz* (pp. 224-229). Félix Alcan.
- Leibniz, G. W. (1990). "Leibniz and Thomas Burnett of Kemney 17./27. Juli 1696" In Akademie der Wissenschaften der DDR (ed.), *Sämtliche Schriften und Briefe* (I, 12, pp. 729-736). Akademie Verlag.
- Leibniz, G. W. (2008). "Essais de Théodicée" In C. I Gerhardt (ed.), *Die philosophischen Schriften*. (Vol. 6). Georg Olms Verlag.
- Fichant, M. (2003). "Leibniz et les machines de la nature", *Studia Leibnitiana* 35, 1-28.
- Mendonça, M. (2017). "La connaissance de la nature: rôle et portée de l'analogie chez La Mettrie" In A. Paschoud & F. Pépin (eds.), *La Mettrie. Philosophie, sciences et art d'écrire* (pp. 279-294). Éditions Matériologiques.
- Plato et al. (2002). *Protagoras*. Paris: Les Belles Lettres.
- Rateau, P. (2022). "L'invention des machines de la nature: physique et théologie", *Revue Roumaine de Philosophie* 66(1), 117-140.

Spaemann, R. (2007). *Persons. The Difference between 'Someone' and 'Something'*. Oxford: Oxford University Press.

Turing, A. (1950). "Computing Machinery and Intelligence". *Mind*, 49, 433-466.

Verbeek, P.-P. (2011). *Moralizing Technology. Understanding and Designing the Morality of Things*. Chicago: The University of Chicago Press.

Chapter 6

Agency in Artificial Intelligence Systems

Parashar Das¹

There is a general concern that present developments in artificial intelligence (AI) research will lead to sentient AI systems, and these may pose an existential threat to humanity. But why cannot sentient AI systems benefit humanity instead? This paper endeavours to put this question in a tractable manner. I ask whether a putative AI system will develop an altruistic or a malicious disposition towards our society, or *what would be the nature of its agency?* Given that AI systems are being developed into formidable problem solvers, we can reasonably expect these systems to preferentially take on conscious aspects of human problem solving. I identify the relevant phenomenal aspects of agency in human problem solving. The functional aspects of conscious agency can be monitored using tools provided by functionalist theories of consciousness. A recent expert report (Butlin et al. 2023) has identified functionalist indicators of agency based on these theories. I show how to use the Integrated Information Theory (IIT) of consciousness, to monitor the phenomenal nature of this agency. If we are able to monitor the agency of AI systems as they develop, then we can dissuade them from becoming a menace to society while encouraging them to be an aid.

¹ Law School, London School of Economics and Political Science -
p.das4@lse.ac.uk.

Keywords: Artificial Intelligence (AI), problem solving, phenomenology of agency, Integrated Information Theory (IIT), indicators of consciousness.

1. Introduction

A prevalent general concern is that the present development in AI systems can result in these systems becoming sentient (Altman 2023; Carlsmith 2023; Future of Life Institute 2023). How can the research community respond to this apprehension; I address this question in this paper. The current generation of AI systems like ChatGPT from OpenAI, LLaMA from Meta, and Google's Bard etc. are developed to be problem solvers. One can feed complex problems to these systems and very often obtain solutions that would otherwise demand a high level of expertise and significant cognitive effort from humans. According to OpenAI, GPT-4 has performed in the top percentiles in academic and professional entry exams (OpenAI, 2023). Researchers at Microsoft have demonstrated GPT-4's capacity to solve novel and difficult tasks in mathematics, coding, vision, medicine, law and psychology. Given the breadth and depth of GPT-4's capabilities, these researchers believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system (Bubeck et al., 2023).

These impressive achievements raise the prospect of AI systems that can surpass human intellectual capabilities in all domains of human activity. This is what philosopher Nick Bostrom calls artificial superintelligence (ASI) (Bostrom, 2014). ASI poses a genuinely novel threat. Bostrom argues that it will not necessarily share common goals with humankind, and may not hesitate to infringe upon human existence in order to realise its own goals. In principle, it will be able to marshal unlimited physical and mental resources to secure ends

fatal to humankind. Moreover much of this may take place in ways beyond our comprehension (Bostrom, 2014).

AI systems develop to be the systems that we forge. To start with, they are trained on knowledge produced by our society. In addition to containing all the greatest achievements in science and humanities, this knowledge is also infected with biases, conflicts, and even neurotic tendencies, to name a few of the malaises permeating our society. These infections get transmitted to AI systems when they are trained. Often the outputs from AI systems highlight these malaises; society uses these outputs and reinforces them; which infect the input to future AI systems, and the vicious cycle continues (Aristodemou, 2014). For instance, deploying the predictive power of AI systems in courts has sparked the concern that demographic and socio-economic based biases in the input data will prompt AI systems to produce biased and discriminatory outputs, thereby reinforcing any judicial tendencies towards racial bias and mass incarceration (Hanna and Bender, 2023; Malek, 2022).

In a recent extensive survey of the capabilities of Large Language Models (LLM) Bowman presents eight potentially surprising claims about LLMs, 'that are reasonably widely shared among the researchers—largely based in private labs—who have been developing these models' (Bowman, 2023). One of these claims is that experts do not understand the inner workings of AI systems. Even though we do not understand the inner workings, it is eminently plausible to argue that, if AI systems do develop threats as foreseen by Bostrom, then this tendency will be acquired from their interactions with our society's negative features. It is important to note that there are no *a priori* properties of *physical information processing systems* that make them inherently subversive to human society. The physical environment is not negatively attuned to human

development rather, as a climate scientist will argue, it is the other way round. It would therefore be equally as hard or easy, to develop an AI system that is a menace, as a system that is an aid, to human development. For example, altruistic ideas can be used to construct AI systems that share and preferentially cherish human goals. These AI systems will be very different from the ones that Bostrom dreads.

In this paper I ask the following question: assuming that AI systems are going to acquire consciousness in the future, how can we monitor the level and the nature of their consciousness as these systems develop? Above I noted that the way AI systems are developed makes them superior problem solvers, first and foremost. I therefore take it as a *working hypothesis* that, the conscious aspects of AI systems would preferentially mirror the conscious aspects of a human problem-solving process. Of course, we can only exploit our knowledge of human consciousness in this investigation. This is a general assumption made by researchers engaged in similar pursuit – for example, in an important survey, Butlin et al. note ‘(a)lthough these theories (of consciousness) are based largely on research on humans.... We claim that using the tools these theories offer us is the best method currently available for assessing whether AI systems are likely to be conscious’ (Butlin et al., 2023). Consciousness in humans has its *functional* aspects – how we behave - and its *phenomenal* aspects – our subjective feelings. An AI system that acquires consciousness will have parallel aspects. I am particularly interested in figuring out whether an AI system as it develops is disposed to be beneficial or detrimental to our society, or what is the nature of its *agency*. The functional aspects of the agency can be monitored using tools provided by the functionalist theories of consciousness. Butlin et al. have made excellent progress in this regard. I concentrate on the phenomenal aspects of the agency. I target the *agentive phenomenology* involved in human problem-solving processes and extrapolate this

understanding to AI systems. I then propose how to use the Integrated Information Theory (IIT) of consciousness, to monitor the phenomenal nature of this agency.

In Section 1, I discuss some cognitive aspects of problem solving. This is necessary to understand the phenomenal aspects of agency involved in problem solving, outlined in Section 2. Section 3 deals with monitoring agency in conscious AI systems. I conclude by observing that monitoring agency will enable us encourage the development of altruistic AI systems while regulating the malicious ones.

2. Some Aspects of Human Cognition in Problem Solving

There are two main approaches to the study of mental processes: cognitive science and phenomenology (Chalmers 1996). Correspondingly, problem solving as a mental process, has its cognitive aspect and its phenomenological aspect. Cognitive analysis provides an account of mind in terms of *what the mind does*. The cognitive features, of problem solving say, are made publicly available by the results of psychological experiments. Cognitive analysis therefore has a *third-person perspective* i.e. a view from the 'outside' (Van Gelder 1999). Phenomenology adopts a *first-person perspective* to pay direct attention to conscious experience i.e. a view from the 'inside'. Conscious experiences, like the experience of problem solving, have their phenomenal character; *there is something it is like* to have those experiences – there is something it is like to think about Pythagoras' theorem, say. Conscious experiences also have their intentional character; they are about the external world, e.g. a problem solver holds a certain *attitude* towards a specific *content* of her problem. Phenomenology deals with both the aspects and brings out *how the experience feels*. To fix upon the agential phenomenology of problem solving, let us begin with the cognitive analysis of a problem-

solving process to better direct our gaze to the relevant experiential qualities.

2.1. A cognitive analysis of human problem solving

A problem occurs when ‘there is an obstacle between a present state and a goal and it is not immediately obvious how to get around the obstacle’ (Goldstein 2018). One has a problem when one intends to achieve a goal and finds the sought after goal has no obvious pathways leading to it. The time development of a problem-solving process involves interactions between employing expertise and thinking creatively. Generalising Stellan Ohlsson’s (2011) insight sequence, the five stages of a problem-solving process are: employing expertise to start with, arriving at an impasse, thinking creatively to overcome the impasse, and employing a creative solution.

At any one stage, a problem solver applies her expertise by drawing upon past knowledge and relevant experiences to analyse the available information regarding the problem situation, to develop an understanding of the goal, and then test familiar strategies to arrive at a solution. However, employing expertise sometimes ceases to generate pathways to a solution. A problem solver becomes stuck, is unable to think of how to proceed, and thus arrives at an impasse. In such cases, creative thinking often breaks the deadlock. The creative problem solver, often generates novel solutions by drawing connections between closely (*near*) or remotely (*far*) *associated* ideas. Usually, a further effort of analysis is required to implement the novel solution or, in the case of a partial solution, further continuation of analysis is required. Either way, there is a return to employing expertise. Problem solving is therefore an interplay between employing expertise and thinking creatively.

i. A cognitive picture of expertise

Experts usually search for the pathway to a solution by applying methods that have previously succeeded in related situations. Gestalt psychologists call this *reproductive thinking*.² Fitting expertise to the problem is a process of systematic analysis, planning, and making incremental adjustments. There is a set of cognitive processes involved in performing analytical tasks that includes: applying the relevant specialist knowledge, recognising patterns in the problem situation by drawing upon a reserve of past experience, weighing between competing pathways and competing goals, employing general reasoning skills, and monitoring progress in order to make necessary adjustments to one's approach (Goldstein 2018).

ii. A cognitive picture of creativity

Creativity is a more involved process. In an influential paper, Sarnoff Mednick (1962) formulated creative thinking in terms of ideas associated with the problem at hand: a problem situation is specified by a set of problem elements and each such element activates specific ideas associated with it. Creativity is the process of combining associated ideas into useful combinations that generate a novel solution. The associated ideas can be closely related to the problem elements, I refer to such ideas as *near associates*, or they can be remotely related or virtually unconnected to the problem elements, and I call them *far associates*. As Mednick puts it, '(t)he more mutually remote the associated ideas of the new combination, the more creative the process or solution' (Mednick, 1962). Following Robert Weisberg

² Gestalt psychologists make the distinction between reproductive thinking and the productive thinking involved in creative problem solving (see Weisberg 2018).

(2018), near associates and far associates can be summarised as follows:

Near Associates: Upon reaching an impasse, a problem solver has exhausted her pool of expertise that is directly related to the problem. The creative problem solver may venture beyond her direct expertise to draw connections with ideas that are near associates. Of course the term *near* refers to a matter of the degree of 'remoteness' i.e. the specificity of the match between the present situation and the associates. To illustrate, a Zika virus researcher may resolve a methodology problem by drawing on her knowledge of past experiments with Ebola virus. This is a near (*analogous*) transfer of expertise from her knowledge of Ebola to the present Zika problem because they share the same conceptual space of a 'virus'.

Far Associates: If no near associated ideas occur readily, then a problem solver must generate novel ideas by drawing links between previously unconnected ideas, these are generally *remotely associated* facts. Here a problem solver's creativity disregards associative connections or mental structures activated in the memory. There are a number of cognitive mechanisms available for this, and relevant for us are the following: *Cognitive disinhibition* increases the range of information available (otherwise cognitively inhibited) to conscious awareness, thus enabling the formation of novel ideas by combining information which are remotely associated (Carson 2014). In a similar vein, a problem solver often focuses attention on the set of relevant ideas (near associates) activated by a problem. When one chooses to pay *broader attention* to a spectrum of ideas that include ideas weakly activated by the problem, another kind of remote association is facilitated (Wegbreit et al. 2012). *Restructuring*, an important mechanism, entails a paradigm shift. When continuing approaches to solve the problem all reach an impasse, and further progress along the

usual way of thinking is foiled by unsurmountable obstacles, it is time to bring about fundamental changes in one's basic understanding of the problem and analytical tactics. To address the crisis the problem solver abandons previous sub-goals, reinterprets the end goal, and endeavours to reach this reformulated end goal, through a novel route consisting of very different sub-goals. In this regard, Kounios & Beeman (2015) make the interesting observation that, remote association and restructuring are not separate processes. Instead, drawing remote associations necessarily involves re-interpreting and restructuring the problem. Here is an example of restructuring which will be useful later.

Robert Buderer (1996) recounts the incident regarding Arnold Wilkins, a physicist working for the Radio Research Station in England in 1935. Wilkins' superior, eager to assist in Britain's war effort, left a memo on his desk: 'Please calculate the amount of radio frequency power which should be radiated to raise the temperature of eight pints of water from 98°F to 105°F at a distance of five km and a height of 1 km.' Wilkins understood that his end goal was to use radio waves to boil the blood of incoming enemy pilots at a distance. However, his calculations showed that the existing technology was incapable of generating such a death ray. Wilkins had reached an impasse. His superior nevertheless insisted; is there any way they could help in the defence of Britain? This prompted Wilkins' creativity. He abandoned the existing end goal of developing a death ray. Widening his focus, he recalled overhearing engineers at the government post office reporting that they had noticed disturbances to their shortwave communications as planes flew by. Wilkins saw the possibility of bouncing radio waves off enemy airplanes to detect them. He *restructured* his problem. The new goal was not to kill enemy pilots but to detect the enemy aircraft sufficiently in advance to mount a defence. He had combined two remotely associated ideas – radio

wave disturbance and locating aircraft – and, thus, had invented the radar (Weisberg 2018).

2.2. Problem solving by AI systems, or an Expertise-Based view of Creativity

I have outlined problem solving as an interaction between analytical procedures and thinking creatively. Can a computing system e.g. an AI system, carry out these cognitive processes? The analytic nature of expertise means that generally, it can be exhaustively described as a rules-based step-by-step procedure, thus constituting a *computable* process. Analytical procedures are therefore amenable to the computing powers of AI systems. But what about creativity?

Weisberg (2018) explains that the most impressive creative solutions can be reached through a continuation of the analytic processes of expertise alone. He stresses that it is impossible to *directly* leap to a remotely connected idea. The problem situation can only activate an idea when a link exists between that idea and the present problem situation. The connections to remote associates are therefore made through a series of near associations, each successive near association building on the preceding one. The most remote connections are built on an existing scaffolding of near associations. Drawing near associations, in turn, involves matching problem-specific or analogous expertise to the problem situation. This, according to Weisberg, is just another way of conducting expert analysis. Additionally, a problem solver may apply heuristics (non-problem specific expertise) i.e. analyse the general logical implications of the problem information (like applying mathematical procedures to the problem, working backwards from the goal, or working forward from the problem information). Hence, drawing both near and far associations are reducible to analytic exercise.

Similarly, solutions to problems that required restructuring can be achieved by analytic thinking alone. Weisberg argues that restructuring can be achieved through a dynamic analytic process. A failure to transfer a past solution to the problem situation can generate new information about the problem. The analysis of that new information prompts a new kind of search of memory in which alternative sets of past solutions and expertise are accessed. This forms the basis for a novel analysis of the problem situation i.e. restructuring the problem. This can lead a problem solver to realise that the initial approach was completely wrong and that an altogether different class of methods is required. The restructured problem, in turn, triggers the retrieval of a new solution type. Furthermore, according to Weisberg, the occasional flash of insight i.e. a Eureka! moment, that delivers an instantaneous solution, is often gained by restructuring the problem situation. A problem solver may repeat several cycles of *restructuring in response to failure* until the insight solution is reached. Hence, even solutions obtained through insight can be achieved through persistent analytic thinking alone.

A dynamic analytic process may remove *any* impasse, in principle. But there is an important caveat, it will suffice only if a problem solver possesses a vast enough reserve of analytical skills and collection of past solutions to make extensive search of their problem field and beyond. Linking far associates through linking near associates alone, often poses an insurmountable obstacle for humans. At any one stage, a problem solver asks: *what next step can I take towards my solution?* For argument's sake, let us say that a problem solver begins with generating 3 candidates near associates. Since these near associates do not solve the problem, he has to look beyond them. For each near associate, looking ahead generates, say, a further 3 possible associates each, thereby increasing the number of associates to 3^2 and, continuing in this manner, will produce 3^3 associates at the next stage and so on.

In other words, if one has to connect remotely associated ideas through a scaffolding of near associations, or what is generally known as a *tree search* procedure, the number of alternatives to be taken into account increases combinatorially as the remoteness increases. Such *combinatorial explosion* simply cannot be handled by human cognition. This is precisely why we rely on creativity instead.

AI systems are very powerful computational systems that can handle combinatorial explosion. For example, the Deep Blue computer chess system, developed at IBM, which defeated grand master Garry Kasparov in 1997, was a massively parallel system designed for carrying out *chess game tree searches*: 'The system (was) composed of ... 480 single-chip chess search engines ... each (single chip) capable of searching 2 to 2.5 million chess positions per second, and communicate with their host node ...' (Campbell et al., 2002). No wonder Deep Blue could conduct an almost exhaustive tree search, evaluate each possible move for effectiveness and thus beat a world champion. In the years since, computer systems have greatly out passed the Deep Blue system, as evidenced by the recent achievements of AI systems like GPT-4 that can rival or surpass human problem-solving efforts.

It would therefore appear that whatever problems humans can solve by applying expertise and creativity, can also be solved by AI systems, through computation. However, these human cognitive processes are equally the locus of the subjective qualities of problem solving associated with them. We do not just solve problems, we feel the difficulties associated and are delighted when our purpose (solution) is achieved. This raises the following question: Will AI systems, as they develop to become superior problem solvers, also develop subjective experience? And if that happens, how can we detect this achieved subjectivity? In particular, I am interested in detecting the

agentive aspect of their subjective experience; will it be their purpose to aid us in solving our problems or hinder, or even subvert, our efforts.

3. Agentive Phenomenology in Human Problem Solving

Following our working hypothesis that conscious aspects of AI systems will mirror the conscious aspects of human problem solving, we need to discuss the agentive phenomenology involved in human problem solving. I will limit myself to considering the phenomenal aspects of agency (see the note at the end of the section regarding intentionality).

3.1. Agency in creativity.

Let us say, I as the problem solver have reached an impasse. I have explored all usual and known avenues towards a solution and have failed. It is now upon me to forge a new path through of my own and accept the consequences whether a success or a failure. There is a *mineness* intrinsic to creative thinking. In agentive action, in general, I have a strong sense of ownership over my action; I feel that I am the one who is performing the action. It involves a sense of self, both as the initiator of action and the experiencer of the effects of the action. The sense of agency, therefore, has a distinctive and proprietary first-person perspective. As Horgan notes ‘... ordinary phenomenology comprises not only the uncontested kinds of phenomenal character, but further kinds as well. It includes self-as-source phenomenology, as an aspect of agentive experience’ (Horgan, 2011). A phenomenal characteristic of agency is therefore a sense of *mineness*; there is something it is like to experiencing one's activity as *one's own actions*.

Dustin Stokes identifies two necessary characteristics of creativity as follows: ‘Some thought (or action) x is minimally creative only if, for

some agent *A*, *x* is the *non-accidental result of the agency* of *A* and *x* is *psychologically (or behaviorally) novel* relative to *A* [my italics]' (Stokes 2014).³ Putting to one side that requirement, what is important here is that, Stokes refers to a non-accidental form of agency, in other words, the sense of agency involved in creative problem solving is distinctly *purposeful*.⁴

In general, agentic action is purposive to varying degrees. Deliberative action is always preceded by a period of reflection in which one weighs between various options until one chooses a particular action for specific reasons. One then performs that action based upon those reasons. This is accompanied by a robust overall what-it-is-likeness of acting *for a specific purpose*, including 'first, the what-it's-like of explicitly entertaining and weighing various considerations favoring various options for action, then the what-it's-like of settling upon a chosen action because of certain reasons favoring it, and then the what-it's-like of performing the action for those very reasons' (Horgan et al., 2003). On the other hand, many everyday tasks involve sequences of actions that are routine and automatic e.g. opening the laptop and checking one's email inbox, or catching a particular bus and getting off at the right stop to reach one's workplace. Despite not being deliberative, these actions nonetheless have a purposive aspect that is subtle and generic: the what-it-is-likeness of acting *on purpose*. Another phenomenal characteristic of agency is therefore a sense of *purposiveness*; the what-it-is-likeness of

³ Here by psychological novelty Stokes refers to 'ideas that are novel relative to some individual mind' (Stokes 2014). This is to be distinguished from historical novelty: 'An idea or act is novel if it is new relative to the history of ideas', as per Boden (2004). Psychological novelty is good enough for our purposes.

⁴ I note that Stokes is considering the conditions for a minimal notion of creativity.

aiming or directing one's actions towards a desired outcome (Mylopoulos 2008).

Now recall the example of Wilkins inventing the radar (Section 1). Applying the usual framework had led to the failure of reaching Wilkins' original goal to develop a 'death ray'. Wilkins' *purposiveness* was triggered when his supervisors insisted on being able to help the war effort. The extent of his purposiveness can be assessed from his willingness to completely abandon his original take on the problem. He could have experimented with increasingly more powerful radio waves, but that would have resulted in a prolonged research programme; he was cognizant of this. He also purposefully disregarded the goal specified by his supervisors and instead of producing a death ray for the pilots, produced a ray for detecting the aircrafts. The sense of agency involved here is rather robust in character, a character that Tim Bayne (2008) calls the 'raw feels' character of agentive experience:

To my mind, the most convincing case for the raw feels account concerns the experience of effort. Consider what it's like to complete a set of press-ups, attempt to solve an unsolvable puzzle, or resist temptation. One's efforts to succeed in these endeavours can be assessed for success, but it is less clear that one's experiences of effort can be similarly assessed.

The purposiveness involved in restructuring a problem is often accompanied by such a 'raw feels' phenomenology. Furthermore, the motivations that drove Wilkins had a strong factor of mineness. He was willing to accept the risks, reap the benefits, and advance his project. Having hit upon his novel solution, Wilkins owned it. He went on to make calculations to show that the energy generated by a one-kilowatt transmitter was sufficient to reflect a detectable level of

energy off the surface of an air plane from five or six miles away (Buderi, 1996).

3.2. Agency in applying expertise

Consider a routine problem-solving process, a problem solver apprehends the problem already pre-figured by her understanding and applies her expertise. There is *purposiveness* in this endeavor, as it is directed towards achieving some level of progress by applying the correct rules and procedures, and the problem solver owns this process reflecting his *mineness*. Thus, implying a sense of agency.

Although there are a few other phenomenal characteristics of agency (Mylopoulos 2008), relevant to my purposes of problem solving are the two I had discussed; purposiveness and mineness. Agency is an ever-present dimension of problem solving. At one pole, there is the 'typical feel' agency of rule application associated with the analytic thinking involved in expertise. At the opposite pole, there is the strong 'raw feels' agency that often accompanies the most creative solutions. And somewhere in between these two poles falls the agentic feeling of any other problem-solving activity; connecting near associates for example.

Note: Regarding the intentionality of agency Horgan et al. write '...this phenomenology (of agency) is richly intentional: it presents oneself, to oneself, as an agent who immanently generates one's own behavior, and who does so in a manner both purposive and voluntary' (Horgan et al., 2003). There is a strong sense of self-presentation involved here. In this paper I disregard the aspects of self-presentation of a conscious AI system, as it can only have second order effects on how it represents us.

4. Monitoring Agency in AI Systems

Like every other mental process *agency* has two aspects to it. A *third-person aspect*, that of studying the cognitive aspects of the way an agent (other than oneself) *functions*; and a *first-person aspect*, the phenomenology of agency as discussed in the last section. Functional theories of consciousness can provide us tools to monitor the functional aspects of conscious agency. To monitor how an agent feels, however, we need a theory that directly approaches phenomenal consciousness, unless of course the agent is willing to self-report.

4.1. Functional aspects of agency

An international collaboration of leading philosophers, neuroscientists, and AI experts, have recently published a report, Butlin et al. (2023), that endeavours to make a scientific assessment of consciousness in AI systems. They adopt *computational functionalism* as their guiding principle. This is the thesis that the brain is essentially a Turing machine and its operations are computations; conscious processes are nothing but computational processes of the right kind. Functionalist theories like the Global Workspace Theory (GWT), Higher Order Theories (HOT), etc. provide accounts of such computations. Using a selection of these theories, Butlin et al. derive a list of indicator properties for assessing consciousness in AI systems. Two of these indicators target agency in AI systems.

Adopting the functional point of view Butlin et al. stipulate, '(s)ystems are *agents* if they pursue goals through interaction with an environment' (Butlin et al. 2023). Their survey suggests that agency is necessary for consciousness as most of the theories they considered make some reference to agency. Some theories require stronger forms of agency, for example, Birch et al. (2020) in their theory of Unlimited Associative Learning, require agents to have flexible goals or values;

while Hurley would require a form of *intentional agency*: ‘...[the system’s] actions depend holistically on relationships between what it perceives and intends, or between what it believes and desires’ (Hurley, 1998).

Agency when present imparts a stronger claim to consciousness than when absent; and intentional agency makes an even stronger claim. Summarizing these considerations they form the following indicator for agency: *Learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals*. They also have an indicator requiring a form of agency with belief-like representations from Higher Order Theories: *Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring*. Here ‘metacognitive monitoring’ refers to the monitoring of conscious states by a higher order representation of that state, as in HOT theories (Carruthers & Genaro, 2023).

It is clear from the nature of the two indicators identified above that they would monitor agency involved in the causal role that agents (AI systems) play. What about the phenomenal aspects of agency then? Functional theories of consciousness leave the analysis of phenomenal consciousness to one side (Van Gulick, 2007; Blackmore & Troscianko, 2018), we therefore need a theory of consciousness that deals with the phenomenal aspects.

4.2. Integrated Information Theory and the phenomenal aspects of agency

The Integrated Information Theory (IIT) of consciousness takes a direct approach to phenomenal consciousness. IIT’s focal point is subjective experience itself rather than its functional properties. It starts by characterizing consciousness through a set of *axioms of*

phenomenal existence. In parallel to these axioms, IIT proposes a set of *postulates for the physical substrate*; these postulates spell out the requirements that must be satisfied by any physical system to support consciousness. In principle, the postulates can be applied to any physical system to determine its degree of consciousness, and more importantly the nature of this consciousness. In contrast to the formalisms discussed in the previous section, IIT does not seek to simulate the computations performed by the conscious brain. We need to discuss the IIT formalism in some detail in order to tease out its implications on agentive phenomenology. The following provides a bare minimum explanation gleaned from the IIT literature (Tononi & Koch, 2015; Albantakis et al., 2023).

The theory is anchored on the premise that for something to exist from the point of view of the world, it must have cause – effect power upon its environment. IIT equates *existence* with *causality*. It now follows that, for something to exist from its own intrinsic perspective, it must have cause – effect power on itself. Consciousness exists from its own *intrinsic perspective* (an axiom), hence for a *physical substrate to support consciousness (PSSC)*, or for any AI system to acquire consciousness, it must have cause – effect power upon itself.

Having made this observation, IIT now proceeds by specifying a systematic manner in which to construct a *cause – effect structure (CES)* of the PSSC that captures its cause – effect power upon itself. As may be guessed, this CES is nothing but a structure capturing the *intrinsic causal flow* within the PSSC; a causal flow from its past intrinsic state, to its present intrinsic state, and on to a future intrinsic state. IIT now postulates the required characteristics for the CES in order that the PSSC may acquire consciousness:

- That conscious experience is *unified* (an axiom), necessitates the irreducibility of the CES; the CES cannot be reduced to a

collection of its sub-structures without severely impeding the intrinsic causal flow. IIT prescribes a positive number Φ that measures the intrinsic irreducibility of the CES; higher values of Φ indicate higher levels of irreducibility.

- Again, each experience is *definite* (an axiom), it has the set of phenomenal distinctions it has, neither less (a subset) nor more (a superset). Quantitatively, this means that the Φ value of the CES, corresponding to the particular experience, is greater than the Φ values of any other CES that has an overlap. This *maximally irreducible cause – effect structure* (MICES) of the PSSC is the winning structure and the value of its irreducibility, being maximal, is denoted by $\Phi^{\max}$.
- IIT has a startlingly precise answer to the question of ‘what consciousness is’: the phenomenal consciousness is identical to the MICES of the corresponding PSSC. This identity should be understood in an explanatory sense: the intrinsic (subjective) feeling of the experience can be explained extrinsically (objectively, i.e., operationally or physically) in terms of cause–effect power (Albantaski 2023).
- Moreover, *the quantity or level of consciousness*—is measured by the $\Phi^{\max}$ value of the MICES. Furthermore IIT provides a definite method for representing (unfolding) the MICES as a constellation of points bound by relations in the cause – effect space. This imparts the MICES a *form (shape)*. Now IIT plays its most intriguing card: The *quality or content of phenomenal consciousness (quale)* is specified by the *form* of the MICES: all phenomenal properties, such as the redness of red, the Eureka! feel of an insight, to mention a few, correspond to different forms of the MICES.
- Finally, the theory provides a calculus to evaluate the $\Phi^{\max}$ value.

4.3. Monitoring phenomenal agency using IIT

Researchers in consciousness over the past decades have developed a plethora of well-developed theories of consciousness. These theories have different underpinnings (Seth & Bayne 2022). For example, in the Higher-order theories (HOT) that we encountered in the previous section, consciousness depends on meta-representations of lower-order mental states. In Global workspace theories (GWTs), consciousness depends on ignition and broadcast within a neuronal global workspace. In IIT, as I had noted before, consciousness is identical to the *MICES* of a physical substrate; IIT is elaborated, explained and argued for in the language of causal flow and this is the *appropriate* language for agency; consequently, IIT develops a concept of consciousness that is anchored on agency.

Let us now go back to agentive phenomenology. From a problem-solving perspective, the relevant phenomenal aspects of agency are *purposiveness* and *mineness* (see Section 2). To explain purposiveness, Mylopoulos (2020) cites David Hume ‘...the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body or new perception of our mind’ (Hume, 2000). Mylopoulos interprets that *giving rise to* is the key phrase that is fundamentally purposive. In IIT formalism, the ‘giving rise to’ is what the intrinsic causal flow achieves; past states of the *PSSC* give rise to future states. Hence the intrinsic causal flow is fundamentally purposive. Since the cause – effect structure determines the causal flow, that structure is fundamentally purposive. Thus, the cause – effect structure or more appropriately the *MICES*, the one which is identified with phenomenal consciousness, captures the phenomenal aspects of *purposiveness*.

I explained (in section 2) the phenomenal characteristic of the sense of *mineness* as follows: there is something it is like to experiencing one's

activity as *one's own actions* – the subjective feeling that I am the author of my own actions. In IIT parlance, my experience of my activities are captured by the intrinsic causal flows. The *MICES* determines all intrinsic causal flows, or it is the author of all such causal flows. I interpret this as the ownership of all causal flow. Hence *MICES* captures the fact that my own activities are experienced as my own. Thus, *MICES* captures both purposiveness and mineness, it therefore captures the phenomenal aspects of agency.⁵ Since conscious experience is identified with the *MICES*, it follows that agency is necessary for conscious experience in the IIT formalism. This is also the conclusion that Butlin et al. draw from their review: ‘One argument for the claim that agency is necessary for consciousness is that this is implied by many scientific theories’ (Butlin et al., 2003).

The measure of the level of consciousness $\Phi^{\max}$, arises from the *MICES*, hence $\Phi^{\max}$ is necessarily a measure of agency and, following the IIT formalism through, I conclude that the *quality* of the agency is specified within the *form* (shape) of the *MICES* when appropriately represented. In principle, these characteristics can be applied to any AI system to monitor the level and the nature of the phenomenal aspects of agency.

Comments: Before I end this section let me note that the formalism of IIT has been implicated with agency in ways that are very different to what I discussed above. For example, Delafield-Butta & Trevarthenb (2022) advance the following point of view: ‘But IIT adapts a linear causality, a “cause-effect” logic ... Consciousness affords a further logic, an “action-effect” causality of experienced self-generated *agent*

⁵ Different authors add other characteristics to agency. See Mylopoulos (2020), for an extensive list. Adding more characteristics would mean a stronger sense of agency. However, as I discuss in the next section, just the character of purposiveness is enough for our objective.

action, animated with feeling and vital purpose [my italics]'. What they mean is that the agent action generates 'action-effect' causality and not the 'cause-effect' causality of IIT. They are, of course, dealing with the third person aspects of agency and I have no disagreement with their proposition. They are silent on the cause-effect causality of IIT, which generates the first-person aspect of agency, its phenomenology, as I hold.

Another view is expressed by Desmond & Huneman, who consider agency and consciousness as duals: 'If agency refers to the "activity" of the organism in relation to the environment, consciousness in its broadest sense denotes the "passivity" of the organism' (Desmond & Huneman 2022). They 'propose that measures of information integration (per IIT) can be more straightforwardly interpreted as measures of agency rather than of consciousness'. I disagree with this proposal; they are considering the third-person aspect of agency here and IIT holds no brief on this matter.

5. Detecting whether or not an AI agent is altruistic or malicious

There are obvious advantages and attendant risks to AI systems developing into superior problem solvers. Philosophers of artificial intelligence have speculated about the behaviours of hostile AI superintelligence. Presently, Bostrom's notion that AI superintelligence will take a 'treacherous turn' dominates (Bostrom, 2014; Carlsmith, 2022, Hendrycks et al., 2023). According to Bostrom (2014):

(B)ehaving nicely while in the box is a convergent instrumental goal for friendly and unfriendly AIs alike. An unfriendly AI of sufficient intelligence realizes that its unfriendly final goals will be best realized if it behaves in a friendly manner initially, so that it will be let out of

the box. It will only start behaving in a way that reveals its unfriendly nature when it no longer matters whether we find out; that is, when the AI is strong enough that human opposition is ineffectual.

From a behavioral standpoint, the malicious AI superintelligence imagined in this vignette is power-seeking, deceptive, manipulative and, above all, supplants any concern for humanity with its own (often anti-human) goals. On the other hand, there are no fundamental principles that prevent us from enabling the development of AI systems that are an aid to society. One would aspire to build altruistic AI systems that cherish human values. Counter to the ‘treacherous turn’, the Center for AI Safety proposes ‘AI assistants [that] could act as advisors, giving us ideal advice and helping us make better decisions according to our own values. In general, AIs would improve social welfare and allow for corrections in cases of error or as human values naturally evolve’ (Hendrycks et al., 2023). Contrasting the power-seeking behaviours of malicious AI agents, altruistic AI agents would behave selflessly, empathetically, cooperatively and, above all, put the needs of humanity above all else. It is therefore vital that we have the ability to monitor the agency, in particular, *purposiveness* of AI systems as they develop, to encourage the ones that cherish our humanness and regulate the ones that despise us.

How can we detect whether a putative AI system is altruistic or hostile? Behavioral analysis of their actions seems to be an obvious answer. To detect conscious agency, Butlin et al. have identified behavioral indicators (Section 3) that involve; goal-oriented learning, responding flexibly to competing goals, and behavior that monitors and updates beliefs in response to the environment. But these agential properties are hardly proprietary to either a hostile or an altruistic AI superintelligence. Similarly, Carlsmith (2022) attributes two cognitive

processes to hostile AI agents. Firstly, *strategic planning*. This refers to the capacity to ‘make and execute plans, in pursuit of objectives, and on the basis of models of the world’. Secondly, *strategic awareness* refers to the capacity to model the ‘causal upshot of gaining and maintaining power over humans and the real-world environment’ (Carlsmith, 2022). Again these are catch-all indicators that correctly target the strategic behaviours of a putative AI system as agentive, but do not distinguish those behaviours that are hostile, from those that are virtuous.

In principle, it is possible for researchers to hone the works cited above and come up with behavioral indicators that can differentiate between good and bad agents, as simply put. This will work if human monitors are not misled. In a recent discussion, Goldstein & Park (2023) argue that a range of existing AI systems have already learned how to deceive humans. The deceptive behavior of the present systems are easy to detect, however AI superintelligences can deceive their developers by initially feigning altruistic behaviours, to be ‘let out of the box’, so to speak. What we need are indicators that inform us, not only of how the agents behave but also of how they feel subjectively.

As discussed in Section 3, IIT directly addresses the phenomenal or subjective feeling of an experience and details it objectively through the $\Phi^{\max}$ value and the *MICES* shape. $\Phi^{\max}$ measures of the level of consciousness, and as I showed, it is necessarily also a measure of agency. This measure can help to specify a scale for risk assessment and regulation, for future conscious AI systems. For example, through experimentation, it would be possible to fix a $\Phi^{\max}$ value above which AI systems tend to gain the capability to use deception, to cheat on safety tests, say, so as to escape human control. At present the European Union is one of the few governmental organizations that specifies a regulatory framework through its AI Act; it assigns each AI

system one of four risk levels: minimal, limited, high and unacceptable (European Commission, 2023). One can similarly use the $\Phi_{\max}$ measure to quantify risk levels, like the above, for future sentient AI systems.

The *quality* of the agency is specified by the *form* (shape) of the *MICES*, when appropriately represented. Each experience corresponds to a distinct shape. Whether an AI system develops to be altruistic or malicious towards humanity can in principle be inferred by fully unfolding the shape of *MICES* of that system. Phenomenally distinct experiences (e.g. agency) are represented by distinct sub-structures within the structure of the *MICES* (see Albantakis et al., 2023). The shapes, corresponding to different qualities of agency, are discovered through observation and manipulation, and these shapes provide the necessary *phenomenal indicators* to monitor the right kind of agency.

Two systems can be equivalent behaviorally, and yet be different architecturally. They will then have different *MICESes* and thus differ phenomenally. Phenomenal indicators, therefore, can detect the true nature of AI systems irrespective of their behavior. A conscious system cannot fake its phenomenology. We know firsthand that feigning altruism and acting altruistically are experiences that differ such that we ourselves do not confuse between them. Part of this phenomenological difference is that feigning altruism involves a *self-interested* purposiveness that is opposite to the *selfless* purposiveness associated with altruism proper. Phenomenal indicators would therefore be able to detect deceptive behavior.

Given an AI system, how to unfold its *MICES* and identify the phenomenal indicators is still a work in progress. An initial attempt in this direction shows how the feel of *spatial* experiences, in particular the feel that space is *extended*, can be accounted for by a corresponding

feature of a *MICES* (Haun & Tononi, 2019).⁶ Although IIT provides a definite procedure to compute the $\Phi^{\max}$ value and compose the shape of the *MICES*, these computations become infeasible when one considers interesting systems that are not outright simple (Mayner et al. 2018). Ideally, one would like to conduct experimental studies to build a *reference library* that matches a particular kind of shape of the *MICES*, or a particular phenomenal indicator, with a particular kind of phenomenal feeling. ‘Indeed, there is much scope for future research to begin mapping psycho- physics, ... onto the geometry of shapes in cause – effect space....’ (Tononi & Koch, 2015). Given the complexity of the brain’s neural architecture this is not yet computationally viable, except for a few initial but encouraging results (Massimini et al. 2010; Calasi et al., 2013). As AI systems develop to acquire consciousness, they will provide relatively simpler architecture to start experimenting and building such a library. With such a reference archive in place, developers can guide and nurture the evolution of an AI system such that it is steered towards achieving altruistic agency, while denying it any scope to achieve malicious agency.

References

- Albantakis et al. (2023) “Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms”, *arXiv preprint 2212.14787*.
<https://doi.org/10.48550/arXiv.2212.14787>.

⁶ The following, as yet unpublished, papers are mentioned in the IIT literature that claim to account for the quality of experienced time [Comolatti et al.], and that of experienced objects [Grasso et al.].

- Altman, S. (2023) "Planning for AGI and Beyond", *Open AI*.
<https://openai.com/blog/planning-for-agi-and-beyond> (accessed 13 October 2023).
- Aristodemou, M. (2014) *Law, Psychoanalysis, Society: Taking the unconscious seriously*. New York: Routledge.
- Birchm J., Ginsburg, S. and Jablonka, E. (2020) "Unlimited Associative Learning and the origins of consciousness: a primer and some predictions", *Biology & Philosophy*, 35(6).
- Blackmore, S. and Troscianko, E. (2018) *Consciousness: An Introduction*. (Third Edition) New York: Routledge.
- Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bowman, S. (2023) "Eight things to know about Large Language Models", *arXiv preprint 2304.00612v1*.
<https://doi.org/10.48550/arXiv.2304.00612>.
- Bubeck, S. et al. (2023) "Sparks of Artificial General Intelligence: Early Experiments with GPT-4", *arXiv preprint 2303.12712v5*.
<https://doi.org/10.48550/arXiv.2303.12712>.
- Buderi, R. (1996) *The invention that changed the world: how a small group of radar pioneers won the Second World War and launched a technological revolution*. New York: Simon & Schuster.
- Butlin, P. et al. (2023) "Consciousness in Artificial Intelligence: Insights from the science of consciousness", *arXiv preprint 2308.08708v3*.
<https://doi.org/10.48550/arXiv.2308.08708>.
- Campbell, M., Hoane, A. and Hsu, F. (2002) "Deep Blue", *Artificial Intelligence*, 134(57).
- Carlsmith, J. (2022) "Is power-seeking AI an existential risk?", *arXiv preprint 2206.13343v1*. <https://doi.org/10.48550/arXiv.2206.13353>.

- Carruthers, P. and Gennaro, R. (2023) "Higher-order theories of consciousness" In Zalta E. and Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2023/entries/consciousness-higher/>.
- Carson, S. (2014) "Cognitive Disinhibition, Creativity, Psychopathology" In Simonton DK (ed.), *The Wiley Handbook of Genius* (pp. 198-221). New York: Wiley Blackwell.
- Casali, A. et al. (2013) "A theoretically based index of consciousness independent of sensory processing and behavior", *Science translational medicine*, 5(198): 198ra105.
- Comolatti, R et al. *Why does time feel flowing?*, In preparation.
- Chalmers, D. (1996) *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.
- Delafeld-butt, J. and Trevarthen, C. (2022) "Consciousness generates agent action", *The behavioral and brain sciences*, 45: e44. <https://doi.org/10.1017/S0150525X2100203X>.
- Desmond, H. and Huneman, P. (2022) "The integrated information theory of agency", *The Behavioral and brain sciences*, 22-24. <https://doi.org/10.1017/S0140525X21002004>.
- European Commission (2023) *Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*.
- Future of Life Institute (2023, March 22nd) *Pause giant AI experiments: An open letter*. <https://futureoflife.org/open-letter/pause-giant-ai-experiments>. (Accessed 27th July 2023).
- Goldstein, E. (2018) *Cognitive psychology: Connecting mind, research and everyday experience* (5th edition), San Francisco: Cengage Learning.
- Goldstein, S. and Park, P. (2023, September 4th) "AI systems have learned to deceive humans. What does that mean for our future?", *The*

- Conversation*. <https://theconversation.com/ai-systems-have-learned-how-to-deceive-humans-what-does-that-mean-for-our-future-212197>.
- Grasso, M. et al. *How do phenomenal objects bind general concepts with particular features?*; In preparation.
- Hanna, A. and Bender, E. (2023, August 12th) *AI causes real harm. Let's focus on that over the end-of-humanity hype*. Scientific American. <https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/> (Accessed 16th September 2023).
- Haun, A. and Tononi, G. (2019) "Why does space feel the way it does? Towards a principled account of spatial experience", *Entropy*, 21(12): 1160.
- Hendrycks, D., Mazeika, M. and Woodside, T. (2023) "An overview of catastrophic AI risks", *arXiv preprint 2306.12001v6*. <https://doi.org/10.48550/arXiv.2306.12001>.
- Horgan, T. (2011) "From agential phenomenology to cognitive phenomenology: a guide for the perplexed" In Bayne T and Montague M (eds.), *Cognitive phenomenology*. New York: Oxford University Press.
- Hume, D. (2000) *A Treatise of Human Nature*. Norton DF and Norton MJ (Eds.), New York: Oxford University Press.
- Hurley, S. (1998) *Consciousness in action*. Cambridge: Harvard University Press.
- Kounios, J. and Beeman, M. (2015) *The eureka factor: Aha moments, creative insight, and the brain*. New York: Random House.
- Malek, M. (2022) "Criminal courts' Artificial Intelligence: the way it reinforces bias and discrimination", *AI and Ethics*, 233-245.
- Massimini, M. et al. (2010) "Cortical reactivity and effective connectivity during REM sleep in humans", *Cognitive neuroscience* 1(3): 176-183. <https://doi.org/10.1080/17588921003731578>.

- Mayner, W. et al. (2018) "PyPhi: a toolbox for integrated information theory", *PLOS Computational Biology* 14(7).
<https://doi.org/10.1371/journal.pcbi.1006343>.
- Mednick, S. (1962) "The associative bias of the creative process", *Psychological Review*, 69, 220.
- Mylopoulos, M. and Shepherd, J. (2020) "The experience of agency" In Kriegel U (Ed.) *The Oxford Handbook of the Philosophy of Consciousness*. New York: Oxford University Press, pp. 164-187.
- Ohlsson, S. (2011) *Deep Learning: How the mind overrides experience*. Cambridge: Cambridge University Press.
- OpenAI (2023) "GPT-4 Technical Report"; *arXiv preprint* 2303.08774.
<https://doi.org/10.48550/arXiv.2303.08774>.
- Seth, A. and Bayne, T. (2022) "Theories of consciousness", *Nature reviews neuroscience*. 439-452.
- Stokes, D. (2014) "The role of imagination in creativity" In Paul, E. and Kaufman, B. (eds.) *The philosophy of creativity: new essays*. New York: Oxford University Press.
- Tononi, G. and Koch C. (2015) "Consciousness: here, there, and everywhere?", *Phil. Transc. R. Soc. B*.
<http://doi.org/10.1098/rstb.2014.0167>.
- Van Gelder, T. (1999) "Wooden iron? Husserlian phenomenology meets cognitive science" In Petitot, J. et al. (eds.) *Naturalizing phenomenology: issues in contemporary phenomenology and cognitive science*. California: Stanford University Press.
- Van Gulick, R. (2007) "Functionalism and Qualia" In Velmans, M. and Schneider, S (eds.) *The Blackwell Companion to Consciousness*. New York: Wiley Blackwell.
- Wegbreit, E. et al. (2012) "Visual attention modulates insight versus analytic solving of verbal problems"; *The Journal of Problem Solving*.

Weisberg, R. (2018) "Expertise and structured imagination in creative thinking: reconsideration of an old question" In Ericsson, K. et al. (rds.) *The Cambridge Handbook of Expertise and Expert Performance* (Second Edition). Cambridge: Cambridge University Press.

Chapter 7

Ethics, Artificial Intelligence and Machine Translation in Literary Translation

Daniela Alexandra Rodrigues¹ and Ana Rita Pina²

In an effort to keep up with the ever-growing globalised world, some book editing companies or self-publishing authors find themselves searching for solutions to speed up the translation process and cut its costs, which might bring up some ethical issues and compromise the quality of the work they put out. Furthermore, the use of Artificial Intelligence (AI) and Machine Translation (MT) raises a set of ethical issues regarding, for example, the use of copyrighted materials, loss of work opportunities for actual professional translators, lower remuneration and recognition of translators' work, the lack of transparency by these companies when it comes to the "authors" of these translations, accountability, data ownership and confidentiality, etc. In this paper, we aim to explore some of the issues that have been arising in the translation industry, namely the literary one, regarding the use of LLMs and MT by publishers.

Keywords: machine translation (MT); artificial intelligence (AI); ethics; Large Language Models (LLMs); translation

¹ Centre for English, Translation and Anglo-Portuguese Studies, Faculdade de Letras da Universidade de Lisboa - d-alexandra@edu.ulisboa.pt.

² Faculdade de Letras da Universidade do Porto - anaritarpina@gmail.com.

1. Introduction

The question of ethics in translation, particularly in translation technologies, has recently started to become more evident, especially with the rise of new translation technologies, such as Neural Machine Translation and Large Language Models (LLMs) that use generative AI, that bring on a new set of issues. The ethical side of technology, however, is highly debated. According to Koskinen & Pokorn (2021, p. 2), ethics comes from the Greek expression “*ta êthika*”, which is related to things pertaining to ethos, in other words, it pertains to one’s actions and character in certain situations, where one’s moral compass can sway in two (or more) different directions. Ethics can also be described as the “science which is concerned with moral behaviour or with right or wrong and good or evil of human behaviour” (Sudan, n.d., p. 3) or a “branch of philosophy that used to study ideal human behaviour and ideal ways of being” (Rich, 2020, p. 4). According to these definitions, ethics can be defined as a set of moral principles, rules and/or values that deal with how individuals ought to conduct themselves in society. Therefore, we can perceive ethics as being a guide so that an individual can ponder over their actions in order to choose the best decision for them (morality) and/or the group (society) and present a behaviour that is beneficial to various individuals - and/or society in which these individuals cohabit.

When it comes to ethics, the translator’s job is often restricted by a set of ethical rules which we call deontology. According to Davis (2000, p. 205-128 *apud* Koskinen & Pokorn, 2021, p. 3), deontology “refers to duty- or rights-based ethics that aims to normatively guide what we ought to do, rather than how we should be or what outcomes we should produce and insists that our lives should be governed by moral rules that should not be broken regardless of the consequences”. The

question of ethics in translation studies can be very wide and affect many different topics. As per Robinson (2003, p.25), translation ethics have been “strictly defined as the practice to keep the meaning of the source text undistorted”. For instance, Chesterman (1997a, 1997b, *apud* Hermes, T., 2009, p. 96) sees ethics as “related norms to professional ethics, which, he claimed, demanded a commitment to adequate expression, the creation of a truthful resemblance between original and translation, the maintenance of trust between the parties involved in the transaction and the minimization of misunderstanding.”. However, with technological evolution came other problems not just related to translational action. Moorkens (2022, p. 22), for example, talks about a general consensus that technology is ethically neutral, but according to the author, many ethicists and philosophers of science defend that it is not neutral, but rather a reflection of their designer. Machine Translation (MT) systems and LLMs rely on training data, which is retrieved from different sources chosen, which then moulds the way the system works, something we see in MT systems that are trained for specific uses.

One other problem which relates to the training data is, for example, the use of copyrighted material. This can even become a bigger problem when we are talking about translations of literary works, which are often not considered original work of the author and thus not interpreted as copyrighted material (Moorkens, 2019, p. 3). Not only translations but even source materials and the original texts have been allegedly used to train these systems by using unauthorised versions of many renowned authors without paying any licensing fees, instead of using public domain works, which has already

resulted in a lawsuit from 18 authors against ChatGPT³⁴. Considering the fact that both NMT and LLMs recycle everything we input into these systems, either for learning or for other objectives, companies and publishers are depositing their copyrighted material of their own volition into these systems, unaware of this recycling.

However, the ethical issues surrounding translation technologies involve much more than only copyrighted material. The lack of transparency when using these new technologies has also become problematic, with cases of publishers using MT and LLMs to translate and sell machine-translated or AI-translated books, as we have seen in Portugal or China. This problem is even bigger when we are referring to post-editing, for example. Should a company or publisher disclose that the translation is not human, but rather Machine Translation Post-Editing (MTPE)? How does the use of MT or LLMs in literature affect its readers' opinions, experiences and perspectives of the book and their respective authors? The use of these technologies can also damage the reputation of the company, of the author and the translator, and even bring more serious risks. According to Canfora and Ottman (2018, *apud* Moorkens and Lewis, 2019, p. 9), the publication of raw MT output carries reputational risks, as well as legal and communicational, which can turn disastrous, causing possible harm to the public, for instance, unusable products or damage of said products, injury or even death.

³ Source: Gerken, T. & McMahon, L. (2023). *Game Of Thrones author sues ChatGPT Owner OpenAI*. Accessed on: November 20, 2023. Available at: <https://www.bbc.com/news/technology-66866577>

⁴ Source: Zhan, J. (2023). *George R.R. Martin among 17 authors suing over ChatGPT*. Accessed on: November 20, 2023. Available at: <https://www.vulture.com/2023/09/george-r-r-martin-authors-guild-chatgpt-lawsuit-openai.html>

We also know that using MT and LLMs in literature can disrupt the literary translators' voice and also deliver a translation influenced by another translation present in the training data (cf. Kenny & Winters, 2020). According to Kenny (2011, p. 5) "SMT [Statistical Machine Translation] systems make decisions, but they do so by using statistical models and algorithms designed to select, as efficiently as possible, the 'best' option from a large number of hypotheses built using those models.". Therefore, these systems and technologies are not able to make ethical and pragmatic - and at times grammatical - choices in all planes.

Another problem relating to MT and LLMs is that many of them are free, easy to use and widely available. This makes it extremely easy for anyone without any MT or AI literacy to use without any knowledge of the risks and consequences of its use. For example, systems like Google Translate, DeepL or ChatGPT retain all input text and reuse it for training purposes which raises confidentiality issues such as the reuse of sensitive and private information, personal information and even commercial information and "Not only that, but people who reuse translations through Google Translate cannot be sure that those translations have been shared by their rightful owners in the first place, and nor can they acknowledge the anonymous translators whose work they are reusing." (Kenny, 2011, p. 6). These systems are often referred to as having a "black box", due to the fact that we don't know what goes on inside the process of the translation, which once again risks the leak of this sensitive information and even the reuse of copyrighted material, as referred previously. This recycling of information becomes a problem not only due to the contents of the texts inserted in these systems, but also the exposure of "metadata such as when the text was translated, by whom, and maybe even from which IP address will be possibly shared amongst the different stakeholders" (Escartín & Moniz, 2019, p. 16).

Issues such as who has access to the data, who is the data curator and manager, how is the data processed and where and how it is stored are key prior to establishing any translation workflow to ensure that all parties are protected from potential data and privacy breaches, or even potential threats like cyberattacks (for instance, all data could be anonymized, whenever possible, and encrypted to avoid data breaches should this happen). (Escartín and Moniz, 2019, p.7)

This also hurts the translators' profession, as most companies are willing to give up on the use of human translators and prefer just sticking to AI since it allows them to cut costs and speed up production. Although these technologies are revolutionary and changing the industry, it does not mean that we should not care about the consequences of its use, especially when it comes to ethical issues. In the following sections, we will expose and discuss these issues further, as well as explore some of the ethical problems that have arisen in recent years relating to the use of LLMs, AI, and MT systems in Literary Translation.

2. Core Questions Within the Use of LLMs and MT in Literary Texts

With the increased popularity and interest in LLMs and MT, there has also been an increased number of literary texts being published using these two technologies. From classic literature to modern web novels, it seems most publishers are betting on these new technologies to speed up translation and cut costs, disregarding completely the risks and ethical issues that they entail. Although both MT and generative AI have come very far, it is still apparent that they still have many limitations. The presence of gender biases, racist and offensive language, as well as other problems are topics that have been extensively studied throughout the years. Another relevant problem

is the limitations these technologies have with low-resource language pairs, which can turn disastrous. However, that has not stopped many publishers and companies from using these technologies to translate books and novels.

Recently, some Chinese web novel platforms have resorted to using AI and/or MT to translate their novels more quickly, and even companies are now offering AI-oriented solutions for promoting content overseas by using AI-assisted translation claiming to be the “first of its kind worldwide” and to be “producing massive multilingual literary works in a highly efficient way at a low cost”⁵⁶. There are allegedly accounts of some Portuguese publishers using AI to translate a wide array of literature, including Russian and English classics, where the same author appeared to translate dozens of literary pieces, such as *War and Peace* and “other novels”⁷, in a short amount of time. There are also cases of books being sold on Amazon with Google Translate or DeepL as their co-authors⁸. APTRAD, the Portuguese Association of Translators and Interpreters, even published a statement on November 2023 exposing the growing cases of MTPE requests for literary classics in Portugal and explaining that

⁵ Source: Shen, X. (2020). Chinese publishers see AI translation as the killer app for pushing web novels overseas. In *South China Morning Post*. Accessed on: October 18, 2023. Available at: <https://www.scmp.com/abacus/tech/article/3102891/chinese-publishers-see-ai-translation-killer-app-pushing-web-novelsand>

⁶ Source: Funstory.ai (n.d.). Accessed on: October 18, 2023. Available at: <https://funstory.ai/en/home-english/>

⁷ Source: Vale, F. (2023). *Traduções por inteligência artificial (IA) chegam a Portugal sem se fazer anunciar*. In *Comunidade Cultura e Arte*. Accessed on October 18, 2023. Available at: <https://comunidadeculturaearte.com/traducoes-por-inteligencia-artificial-ia-chegam-a-portugal-sem-se-fazer-anunciar>

⁸ Source: Joan Russel. LA TRAVIESO MONJA: Freira-Fomania (Portuguese Edition). Accessed on: 18/10/2023 Available at: https://www.amazon.it/gp/product/B00KO4ZVN8/ref=dbs_a_def_rwt_hsch_vapi_tkin_p1_i3

“machine translation systems are not able to recognize and transpose all elements of literary works” and that companies and publishers are resorting to these systems to do a two-man job for the lowest price possible⁹.

It is clear that the use of these new technologies is becoming more evident, however, not many seem to understand the risks and ethical problems that come with it. The use of MT, AI and LLMs in literary translation seems to be growing, but the ethical side of selling and publishing books translated by MT or AI is questionable. As a result, the number of ethical issues resulting from this (mis)use have been growing, with many researchers looking at potential problems related to MT and AI use in literary and creative texts. We expose in the next sub-chapters some of the core questions that have been researched so far and briefly summarise them based on the state of the art.

2.1 Transparency, accountability, and regulation

One problem that seems to be becoming more relevant, however, is the lack of transparency regarding the use of these new technologies, which may lead to numerous problems. The lack of transparency when using these technologies could not only impact the relationship between publishers and/or companies and their clients, but could also affect their reputations (Canfora and Ottman, 2018, *apud* Moorkens and Lewis, 2019, p. 9). As stated by the European Parliamentary Research Service (2020, p. 29) “The overwhelming consensus amongst the research community is that trust in AI can only be attained by fairness, transparency, accountability and regulation.”¹⁰. Another

⁹ Source: APTRAD. (n.d.) *O uso de inteligência artificial no meio editorial*. In APTRAD. Accessed on: January 17, 2024. Available at: <https://aptrad.pt/o-uso-de-inteligencia-artificial-no-meio-editorial/>

¹⁰ Source: European Parliamentary Research Service. (2020). *The ethics of artificial intelligence: Issues and initiatives*. Accessed on: December 12, 2023. Available at:

possible issue regarding transparency and trust in the use of MT and LLMs in literary translation revolves around the idea that consumers believe the translations they are buying are human, which is the norm, and usually implies that the text has passed a revision/proofreading, quality assurance, and other controls. One could argue that this question could be easily solved by doing machine translation post-editing, however, this involves more time and a higher cost for the companies and publishers and can also affect the quality and put the authors' work at risk. By allowing the distribution of non-controlled literary works, we are flipping a coin in terms of quality and sales.

Recently, the European Council of Literary Translators' Association (2023, p. 1) released a statement in which they demanded the imposition of transparency requirements on the use of AI to be properly enforced and urged them to clearly disclose its use on the final product when or if they used AI, regardless of the number of times or at what point it was used, appealing to each participant in the production chain to be accountable for their use of AI and to be fairly compensated for their specific work. The statement demonstrates the importance of regulation and control of the use of AI, LLMs and MT, and reflects the state in which the (literary) translation industry finds itself. The lack of transparency, accountability and regulation can bring up a number of questions regarding not only the end product, but also other questions related to data ownership, privacy and the impact on the translation industry.

There is also the question of mistakes, errors and other controversies that might arise from fallacies in the translation, such as a lack of cultural awareness, use of offensive language, for instance, possible racist, xenophobic, homophobic or transphobic translations, as well as

any other biases, and how we cannot determine who is accountable for the final translation. Most of the time we cannot even explain the AI's decisions due to their black boxes' untraceable decision-making, which is why, for example:

The AI Now Institute at New York University, which researches the social impact of AI, recently released a report which urged public agencies responsible for criminal justice, healthcare, welfare and education to ban black box AIs because their decisions cannot be explained. The report also recommended that AIs should pass pre-release trials and be monitored 'in the wild' so that biases and other faults are swiftly corrected. (European Parliamentary Research Service, 2020, p.32 *apud* AI Now Report, 2018)

So, when AI or MT deliver problematic translations, there should exist regulation which holds someone accountable "whether that be the designer, the developer or the corporation selling the AI" (European Parliamentary Research Service, 2020, p.46). Many times, the author of the original source text might not even be aware that the publisher is using AI or MT to translate their intellectual property, and this can result in complicated problems when we are trying to determine who is accountable for the final translation. According to Novelli et al (2023, p. 5) , "Various factors lead to (uses of) AIs perpetrating wrongdoings, consider for example the case of AI perpetrating undue discrimination, this can result from biased training data, system bugs, programmer errors, misuses, or the replication of social discrimination; and sometimes a combination of these factors." but "Addressing accountability in AI requires tackling several difficulties, like the broad definition of accountability and the opacity of AIs." (p. 11). The EU has also recently released the Artificial Intelligence Act which sets out the following:

(...) generative foundation AI models (such as ChatGPT) that use large language models (LLMs) to generate art, music and other content would be subject to stringent transparency obligations. Providers of such models and of generative content would have to disclose that the content was generated by AI not by humans, train and design their models to prevent the generation of illegal content and publish information on the use of training data protected under copyright law. - (European Parliamentary Research Service, 2023, p. 10)

However, we know that this is something extremely hard to control. In the same document, we do not seem to find any description of what a “low-risk” or “minimal-risk” use of generative AI and LLMs could be, leaving it ambiguous on what the obligations are, writing “All other AI systems presenting only low or minimal risk could be developed and used in the EU without conforming to any additional legal obligations.”. The question here pertains to the fact that the use of these technologies should be adequately regulated and their risk level defined, as cited in the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) “Stakeholders also highlighted that besides the clarification of the term of AI, it is important to define ‘risk’, ‘high-risk’, ‘low-risk’, ‘remote biometric identification’ and ‘harm’.” (2021, p.8). Otherwise, until then, we are allowing companies and/or publishers to use AI, LLMs and MT indiscriminately without any type of regulation, transparency or accountability measures by not establishing regulation on the use of MT and AI for official and professional translations.

The need for accountability, regulation and transparency is clear, with many translators’ associations releasing statements demanding the use of MT, LLMs, or AI in general, to be disclosed in the translations and asking for regulation and transparency rules for the use of these

new technologies. Not only the European Council of Literary Translators' Association, as mentioned before, but also APTRAD (the Portuguese Association of Translators and Interpreters), as well as ATRAE (Spanish Translation and Audiovisual Adaptation Association)¹¹, stating "(...) ATRAE, se posiciona en contra de la posesición e insta a las grandes plataformas audiovisuales a que destierren estas prácticas y dejen de perjudicar a un sector ya demasiado precarizado", and ATLAS (Association for the Promotion of Literary Translation), who joined ATFL (Literary Translators Association of France), to denounce the situation of literary translation in France, appealing for more transparency and for regulation that protects the translators' profession: "our two organisations issue a public warning about the industrialisation of this technology and call for immediate transparency on the part of book professionals about these practices which also generate a considerable amount of digital pollution.", "We call for transparency" and "We must demand that the publisher provide information on how a translation has been carried out in the case of 'pre-translation' by AI"¹².

Considering the growing cases of MTPE or raw MT output use, we can determine that this is not only affecting the translation industry, but also having negative outcomes for both the authors and the readers, which in turn makes readers stop purchasing products, and affects the author's in monetary and reputational terms. There are many other problematic questions that also arise from the lack of transparency, regulation and accountability, however. It is evident

¹¹ Source: ATRAE. Comunicado sobre la posesición. Accessed on November 25, 2023. Available at: <https://atrae.org/comunicado-sobre-la-posesicion/>

¹² Source: Association pour la traduction littéraire, (2023) IA and literary translation: translators call for transparency. Accessed on November 25, 2023. Available at: https://atlf.org/wp-content/uploads/2023/05/ENG_AI_and_literary_translation__Trad_Shaun_White_side.pdf

that there needs to be a change in the way we handle these issues in the use of AI, LLMs and MT in (literary) translation.

2.2. Data Ownership and Confidentiality

Another big issue that arises in regard to ethics in AI, LLMs and MT relates to data ownership and confidentiality of copyrighted material. According to Kamocki and O'Regan (2016, p. 4461), the use of free online MT may entail privacy risks which many users are unaware of and that MT service providers may be tempted to take advantage of. We consider that this is also applicable to the more modern LLM and AI tools widely and easily available, taking into account the fact that they work in similar ways. As of 2021 alone, Google Translate has recorded more than 1 billion users¹³, in 2023, DeepL records more than 10 million users monthly, of which 500 000 pay for the premium subscription¹⁴, in 2023, ChatGPT has recorded more than 1.5 billion users, receiving more than 100 million users per day¹⁵ and Google Bard registered 146.6 million monthly visitors. Bowker (2020, p. 266)

¹³Source: Shankand. S. (2013). *Google Translate now serves 200 million people daily*. In CNET. Accessed on: November 25, 2023. Available at :<https://www.cnet.com/tech/services-and-software/google-translate-now-serves-200-million-people-daily/>

Pitman, J. (2021). *Google Translate: One billion installs, one billion stories*. Accessed on: November 25, 2023. Available at: <https://blog.google/products/translate/one-billion-installs/>

¹⁴ Source: Shrivastava , R. (2023). *This German Unicorn Is Trying To Take On Google Translate And ChatGPT*. In Forbes. Accessed on: November 25, 2023. Available at: <https://www.forbes.com/sites/rashishrivastava/2023/08/08/deep-l-is-trying-to-take-on-google-translate-and-chatgpt/>

¹⁵ Source: Mahajan, V. (2023). *100+ incredible ChatGPT statistics & facts in 2024*. In Notta. Accessed on: November 25, 2023. Available at: <https://www.notta.ai/en/blog/chatgpt-statistics#:~:text=Key%20ChatGPT%20Statistics,-If%20you're&text=ChatGPT%20currently%20has%20100%2B%20million,GPT%2D4%20from%20March%202023>

explains how some users of these MT systems may believe that once you have finished your translation on the MT system, the text simply disappears forever. However, we know that this could not be further from the truth.

The most common practice is for the input of these MT and AI systems to be “recycled” for training (e.g. training data). Although citizens’ data might be protected by law and regulations, such as the new General Data Protection Regulations (GDPR), Kamocki and O’Regan (2016, p. 4461) conclude that “It should be openly admitted that an online MT service that fully complies with the Directive is possible only theoretically.”, especially when regulations only apply to personal data, not the ‘anonymous’ data that are usually used to train models. In other words, users should not expect these MT and LLM providers to comply with the law and take precautionary measures to prevent their data from being stolen. Furthermore, personal data or any other information about individuals who were unaware of being involved in the training process can be restored from a model in a way that might significantly affect the regulation of said systems.

But despite people having rights about how their personal data is used and stored, they still have very few rights over trained models (European Parliamentary Research Service, 2020, p. 13). According to the European Parliamentary Research Service (2020), this raises a number of ethical questions regarding, for instance, the level of control subjects have over the data that is collected about them, an individual’s rights to use the model or to know its purpose, taking into account their part in its training. But it also takes into account questions relating to how these systems could inadvertently violate people’s privacy, and how to prevent individuals involved in training a model identity and personal information from being discovered, for instance through potential cyber-attacks. The amount of inadvertently

disclosed sensitive information and possible confidential information breaches becomes even more worrisome, especially if we consider the fact that these systems are developing, expanding, and broadening their availability more each day.

Users are also often misinformed when it comes to MT and AI, and the majority of them do not know how to use these technologies in a critical way (Bowker, 2021, p. 26). This can become problematic due to the fact that these systems are free and widely available and have millions of users each day. For example, many users are unaware of the fact that the information that they input into these systems does not get erased from the system even after they stop using it or that these systems use this data to train the models and store it for other purposes. This raises other questions, for instance, who has access to the data, who is the data curator and manager, how the data is processed and where and how it is stored are essential prior to establishing any translation workflow in order to guarantee the data and privacy safety of every user (Escartín and Moniz, 2019, p. 16). If we keep in mind the number of individuals that use these systems daily by inputting more sensitive information, such as in healthcare, legal, crisis or other bureaucratic situations, where their personal information might be disclosed and exposed in these systems only to be reused for the training of MT or other purposes, possibly without any curating and their consent.

The question of data ownership has also recently been brought up by some researchers. For instance, Moorkens (2022, p. 124) explains how MT training data is stored in parallel or aligned bilingual segments of texts translated by humans and how this translation data is likely to be shared in public repositories such as the European Commission's Directorate-General for Translation data, which can then be repurposed and disseminated for free, in both commercial and non-

commercial settings, privately held repositories of translations data or parallel data taken from the web. According to Moorkens and Lewis (2019, p. 125), “big translation data has become a valuable resource for MT and machine learning system training”, but translators do not get any extra remuneration for it. Taking into account the lack of details of said data, it means the source of the training data, and its translators, are usually not identifiable.

Data privacy and confidentiality, as well as data ownership, are therefore some of the primary ethical questions brought up by the use of LLMs and MT. These questions can become particularly tricky, being even considered “one of the most puzzling problems in copyright law” (Wu, 1997, p. 154 *apud* Moorkens & Lewis, 2019, p. 8), especially if we take into account the fact that, for privacy reasons, the metadata that identifies the author of the translation, as well as other details, are erased prior to MT training. We can see this situation as having a both a positive and negative side, where on one hand we are protecting the information of the users, but on the other, we are erasing the identity of the translator and their right to the ownership of the data, particularly when we are talking about enormous amounts of data, such as the case of LLMs and MT systems, which make it impossible to search for the author of a translation. According to Moorkens & Lewis (2019, p.7), “an attempt to unpick copyright ownership and intellectual contribution would hinder research progress for MT contrary to the intention of copyright law”.

For instance, in Europe, laws are apparently able to provide only some limited data protection. The current intellectual property laws are not enough to establish a satisfactory framework for managing data since each country has its own scope and extent of each right and regulation regarding data ownership and privacy: “For example for the same fact pattern copyright or trademarks may be enforceable in one country

but not another" (European Commission. Directorate General for Communications Networks, Content and Technology & Osborne Clarke LLP., 2016, p. 12).

Copyrighting AI or MT-generated material is also dubious, because not only are the laws and regulations different all over the world, but also it is not clear how we should attribute copyright to a conglomerate of AI-generated text based on other people's intellectual property. Moorkens & Lewis (2019, p.8) give us two examples of laws that try to reach a consensus on how we should copyright these materials. The first example is the UK Government Copyright, Designs and Patent Act (1988), which states that "in the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken". The second is the final report of the US National Commission on New Technological Uses of Copyrighted Works (CONTU; 1979) which sets out that the 'author is the one who employs the computer'. According to Cabanellas (2015, *apud* Moorkens & Lewis, 2019, p. 9), raw MT output will not be subject to copyright.

Data ownership and privacy are becoming an even bigger issue when we look at how valuable language data is becoming. With the booming development of technology since the 2010s, AI and NMT require more and better training data to develop their systems in order to provide higher-quality output. This data is being sold for machine learning system training at high values, with reports of even billion-dollar valuation, resulting in a "language data boom" (Diño, 2018). The majority of the higher quality training data, however, is human data, which is often obtained through ambiguous methods, such as web crawling (Toral et al, 2017, *apud* Moorkens & Lewis, 2019, p. 2)

and aligned human translations, leading to a problem in regard to the privacy of individuals.

The question of data ownership and confidentiality is still a big issue that requires more thought and planning, as well as a need to be duly established under the different laws and regulations - namely European law, and its various member states, particularly when it comes to MT and AI and translation. Not only can the lack thereof affect translators' and authors' data, but also other individuals' information.

2.3 Impact on the Translation Industry

As the interest in the use of MT and other AI systems in translation seems to be growing, the impact that it creates on the industry seems to bring more harm than good. There seems to be a growing consensus that AI will have a major impact in many areas previously considered to be immune from this kind of technology. Translation and translators alike are often disregarded. However, translation is a highly skilled task that requires more than just knowing a couple of languages. According to Venuti (1995, p.1), the more a translation presents a higher quality, the less visible are the marks that it is a translation, making it so the author of the original source text stands out, even in a foreign language.

when it reads fluently, when the absence of any linguistic or stylistic peculiarities makes it seem transparent, giving the appearance that it reflects the foreign writer's personality or intention or the essential meaning of the foreign text—the appearance, in other words, that the translation is not in fact a translation, but the “original.” (Venuti, 1995, p. 1)

The ever-growing use of these systems makes it so that companies and publishers feel the need to dispense of these skilled individuals, as the costs of using an MT are significantly lower than keeping a certain amount of human translators around. And many other problems are underway. Since around 2010, the market for MT post-editing has been expanding, even before the use of NMT. Instead of using MT and AI to the translators' advantage, making it so they can more easily or quickly do their job, they are instead being used as a tool to replace these skilled workers. There are translators who have witnessed their work being shifted to the post-editing of often low-quality MT output, which might not necessarily be a problem for everyone. Yet, most translators oppose this type of task, going as far as saying they would prefer leaving the industry than resorting to MTPE. However, what often happens is that some companies and/or publishers are even skipping this step to cut costs, a problem that is becoming more problematic when we look at the quality of the work that is being published. It is a fact that some translators enjoy this type of post-editing work, however, the majority tends to prefer doing more meaningful work, meaning that they would prefer to better their craft while being motivated by a sense of achievement and recognition by the readers.

If this is not a consideration in translation production, better workers will leave and there will be a shortage of skilled translators and/or post-editors. Such a shortage would affect reliable access to multilingual information and gathering high-quality bilingual data on which MT training relies. (Moorkens, 2022, p. 145).

Dochetyeal (2008, *apud* Moorkens., 2022, p. 133) believes that a suitable and sustainable work system must satisfy the needs of many rather than the few. In addition to that, instead of only focusing on short-term successes, along the lines of productivity and profitability, these

publishers and companies should focus on long-term effects and dynamic efficiencies, such as learning and innovation. Companies and publishers should instead be focused on integrating these types of new technology in the work environment of translators, in order to help the translators' workflow, and the quality of their work, and make it so that it is easier and faster to meet deadlines, instead of trying to simply eradicate the profession.

As mentioned before, the ATLAS/ATFL statement released earlier this year where they appeal for the regulation of the use of AI and MT in the literary translation industry, but they also reveal how this use is hurting the translators' profession, referring "the risk of massively alienating translation professionals" and expose how "AI is not 'intelligent', it imitates human behaviour by plundering what humans have created.". In this statement, both associations also refer to a problem that has been around in the translation industry for the last years: the falling of incomes for translators. With many companies now resorting to MTPE, MT and LLMs or AI to cut costs, many translators find themselves in such hard situations that they feel the need to accept these precarious work conditions: "So why use DeepL and Google Translate rather than have a text translated directly by someone employed to do so? To reduce costs and deadlines, purely and simply." (Association pour la traduction littéraire, 2023, p. 5).

The conditions of the translation industry are already deteriorating with the replacement of translators for AI and MT. The literary translation industry is known as being extremely difficult, with already low rates and stressful environments and deadlines, but it seems to already be feeling the impact of these technological advancements, with royalties gradually dropping. They top it off by asking "Under these conditions, do we really want to delegate our thoughts and skills not just to 'machines', but to those who profit from

them at our expense?“. If the translators’ work continues being undervalued, we are at risk of losing rich cultural and literary treasures and reducing them to a mere conglomerate of data-generated text. Technology should aid the translators, not replace them.

Creative translation should be considered as a national and international literary treasure and deserves to be protected as such. The scraping practices employed for AI training, simply for the purposes of financial gain, violate not only copyright laws but humanity’s cultural heritage as a whole. (European Council of Literary Translators’ Association, 2023, p.2)

With the lack of transparency and regulation, the translators’ profession will continue degrading and the use of unregulated AI and MT might change and damage the profession forever. Although we do understand the necessity for faster and cheaper translations, however, a human being can not rush through various amounts of projects at a time, it needs time to make sure the job is done correctly, otherwise, we risk ending up with low-quality outputs. But publishers and companies using AI and MT to translate and publish material also have to keep in mind the ethical issues and impacts that this use has on the industry. Translating, especially creative material, does not only mean transposing what is present in the original, especially when we do this based on already existing data, which often results in the famous “translationese”. The translational process of creative material is incredibly complex, hence why it takes so long to reach certain solutions.

Every step in the translation process—from the selection of foreign texts to the implementation of translation strategies to the editing, reviewing, and reading of translations—is mediated by the diverse cultural values that circulate in the target language, always in some

hierarchical order. The translator, who works with varying degrees of calculation, under continuous self-monitoring and often with active consultation of cultural rules and resources (from dictionaries and grammars to other texts, translation strategies, and translations, both canonical and marginal), may submit to or resist dominant values in the target language, with either course of action susceptible to ongoing redirection. (Venuti, 1995, p. 308)

The importance of translation, and in this case of the literary industry, is still overlooked. According to Blanchot (1990, *apud* Venuti, 1995, p. 308) “Even “classical masterpieces, live only in translation”. The idea behind Blanchot’s quote is still applicable today, so why are we inputting rich literary works into systems which do not have the capacity to think critically and creatively, as a translator would, ending up putting more low-quality output onto the market at a faster rate, all just to save some money? Translators ought to fight for the rights to their intellectual property - meaning their translated texts - when negotiating the terms of their contracts. They should demand that the texts translated by them should be seen as their own property and, therefore, be titled as such, making sure they receive fair wages for their work and that they receive all the royalties they are entitled to. Furthermore, with the rise of the use of AI and MT, translators should fight for their careers and works to be protected by law, especially when it comes to jobs that use of MT and AI as their “primary translator” to cut costs - which often can include translators’ and authors’ intellectual property already present in the AI’s database.

To summarise, being that the literary translation profession has already been one of the most undervalued, it seems that making these new AI and MT systems more widely available, and for the most part for free, has aggravated the points mentioned previously. The literary translation industry seems to be suffering, as reported by many

translation associations, and ask to be respected as professionals and for the work that they do, as well as for better conditions, regulation and protection of the literary translation industry and of the role of the literary translator.

2.4. Quality

As AI was created by humans, this means that this technology can be susceptible to biases. Due to the data used to train these systems, or due to the values held by the systems' developers and its users, systematic biases may arise. It usually occurs when machine learning applications are trained on data that only reflects the values and characteristics of a small demographic, which normally reflect societal biases. Unless developers work to restructure and counteract these biases, AI systems and products may lead to a perpetual unfairness and discrimination, which can demonstrate extensive effects within certain groups in society. This problem is worsened by the fact that AI systems are usually "black boxes", where it is impossible for the consumer to evaluate whether the data used to train them is fair or representative of their values, making it hard to find these biases. As a result, there has recently been more studies on how to make machine learning fair, less biased, accountable and transparent.

It is known that most MT systems' output is evaluated during training and again afterwards for a quick, easy and cost-effective control of quality. Based on these evaluations, "output may be considered to have reached parity with human translation quality if a segment-level crowdrating or automatic evaluation achieves the same score as a "reference" human-translation target text." (Moorkens, 2022, p. 128). But according to the author this can be problematic, especially when considering research publications and marketing materials, that in turn might be reported, giving it the impression that MT produces

perfect quality output without needing any sort of human translators' input, as if they were not necessary.

Freitag et al. (2021) found that expert (professional translator) evaluators produced markedly different results to crowd workers when carrying out a detailed error analysis with access to full source and target documents, and demonstrated a clear preference for human rather than MT output. (Moorkens, 2022, p. 129)

The subject of quality has been highly debated by many authors. After Besacier's pilot study on the use of MTPE to translate a literary work from English into French, it was observed that common readers did not find anything unusual about the final output, but when the same output was shown to the author's usual translator, he was able to point several mistakes when it came to the style and cultural references. (Besacier, 2014, *apud* Taivalkoski-Shilov, 2019, p. 693). Besacier (2014, *apud* Taivalkoski-Shilov, 2019, p. 693) points out that MTPE can save money and time, but delivers worse quality results. However, he also admits that both authors and readers alike can benefit from MTPE, where readers are able to get access to the work of their favourite authors more quickly (although with lesser quality), and authors can reach wider audiences by translating into more languages.

Taivalkoski-Shilov (2019, p. 694) defends that MTPE is not suitable for literary translation, due to the fact that these types of systems are incapable of understanding texts as a whole, being only able to translate segment by segment, which often alters the meaning and the structure of the source and target texts. Although AI and MT have come a long way, many still believe that they are still not able to fully comprehend and digest literary works as a human would. Guerberof - Arenas e Toral (2022, p. 202) also found that MT literary translation is less creative when compared to both their post-edited or human

translated versions, especially given the bigger amount of errors and less creative shifts, although even MTPE resulted in less creative literary translation when compared to human translation. This is something even the translators have admitted to, revealing that “(...) they find the PE [post-editing] task limits their creativity, and is technically and cognitively more demanding.” (*ibid*, p. 197). In Guerberof -Arenas’ experiment, the readers found the MT translated text challenging, they also pointed out that the MTPE was “not terrible”, but added that the text had “many inconsistencies and the Dutch is not really Dutch it’s an in-between Dutch” (*ibid*, p. 205), which takes us back to the idea of “translationese”, where the translation might be readable, but it does not sound natural.

Some authors argue that the low-quality output in literary translation can do more harm than good. For instance, in literary translation one of the possible poor outcomes is the intelligibility and enjoyability of translated works (see Office of the United Nations High Commissioner for Human Rights, *apud* Taivalkoski-Shilov, 2019, p. 693), as well as “ungrammatical or unidiomatic language on readers’ linguistic skills in the target language, especially in the case of children or young-adult literature” (*ibid.*), but can also interfere in the transfer of literary ideas from one culture to another. Taivalkoski-Shilov considers that, as a consequence, lower-quality literary work, either caused by humans or technology, is a breach of an authors’ moral rights, taking into account the fact that it can be prejudicial to their honour or reputation “and may ruin and ‘decanonize’ their work in the target language (Wang & Domínguez, 2016, p. 301 *apud* Taivalkoski-Shilov, 2019, p. 693).

The matter of quality is therefore vital in literary translation, as well as the need for having a human in the loop for control, an idea also brought up in the European Parliament’s “The ethics of artificial

intelligence: Issues and initiatives" (2020, p. 35) taking into account the fact that "The human can identify misbehaviour by an otherwise autonomous system, and take corrective action" and that "Keeping humans in the loop would also provide accountability" (*ibid.*).

This only confirms the experiments carried out by Guerberof -Arenas and Toral (2020 & 2022), and Besacier (2014) where we observe that in order to achieve a more creative literary translation, with fewer errors and mistakes all across the board, there needs to be human intervention. The need for quality assurance and human intervention is necessary to achieve higher-quality translations, with more research confirming this need, for example, Vardaro et al (2019) explored the quality of NMT and professional post-editing, writing "Results show that quality assurance measures by post-editors and revisors at the DGT are most often necessary for lexical errors."

Recently, a study carried out by Google has shown that MT tends to be more conservative than human translation. The study concluded that "(...) MT is less morphosyntactic diverse and more convergent than HT. We further attribute this difference to the use of beam search that biases MT outputs towards less diverse and less divergent patterns." (Luo et al, 2024, p. 10). The study then concludes by saying that "(...) we show that the presence of divergent patterns in HT has overall an adverse effect on MT quality" (*ibid.*). Many studies have shown over and over that MT and AI translation solutions might be useful, but not for professional use, especially when there is no language professional to supervise and correct possible mistakes. Although MT and AI can deliver good translations in certain contexts, it might not be the best option when you want to translate creative text, such as literary texts, audiovisual, game localization and others.

Although MT and AI output for translation has seen an improvement in quality, it is still known and proven that these technologies do not

hold yet the capability to deliver translations with the same quality, creativity, accurate transfer of meaning, as well as other competencies as human translation. This is especially relevant when it comes to literary text, which involves being able to create creative translations. We can understand, however, that in some cases, companies and publishers and authors alike, might feel the need to use these kinds of AI and MT technologies, but consumers, translators and authors alike should expect some transparency, since it has been proven that AI and MT is still incapable produce the same results as human translation, which, as we have seen in this section, has been proven too shown better results in terms of quality and lack of fallacies.

3. Conclusion

To sum up everything that has been stated so far, as an ever-growing trend, the use of LLMs and MT in the field of translation, and particularly in the literary translation industry, is becoming more and more prominent, resulting in a need for the creation of regulations that ought to control and try to lessen the impact of their use. The use of this recent technology brings up a variety of ethical questions, which impact not only the translation industry, but also the different authors that work in the field. In recent years, it has become more evident that corporations, meaning publishers and editorial companies, as well as self-published authors, have published numerous books using AI and MT to translate literary works in order to “cut costs” and “speed up” the translation process. However, as previously mentioned, they don’t typically meet the requirements of what would be considered a good translation job, as steps to ensure good quality are often overlooked. The need for transparency and regulation is also clear, with many authors and now translation associations appealing for changes in those sectors, claiming that literary works that use MT or AI in any way should disclose it in the final product. This lack of transparency

and regulation can affect authors, translators, publishers and even consumers. Data ownership and confidentiality is still another problem that seems to be far from being resolved, including problems such as lack of MT and AI literacy when it comes to the reuse of data, copyrights, high value of data, etc. The impact of these issues on the industry is enormous, with translation jobs being fewer and less well-paid, making translators having to resort to accepting MTPE jobs that do not pay well and alienate the profession and industry. The quality of AI and MT output is still far from being good enough, with many studies proving that readers prefer human translation for being more creative and that MT and AI output is more conservative and still has many problems with grammar, syntax, lexical choices and other aspects. Therefore, we believe that many of the questions mentioned in this article will still be prevalent and still need to be addressed. We also acknowledge that there are many other ethical questions regarding the use of AI and MT in translation that need further reassessment and research.

References

- Archer, H., Writer-Davies, R., & McGeoghegan, M. (2017). *AI, Automation, and Corporate Reputation*. Ipsos.
- Beauchamp, T. L. (2005). "The Nature of Applied Ethics" In R. G. Frey & C. Heath Wellman (Eds.), *A Companion to Applied Ethics* (1.^a ed). New York: Wiley.
- Bitter Days. (n.d) *Chief Honey Pet: Raise A Little Warm Wife From A Famous Family*. Webnovel. Accessed on: January 30 2024. Available at: https://www.webnovel.com/book/chief-honey-pet-raise-a-little-warm-wife-from-a-famous-family_15659766605558705

- Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., & Ueffing, N. (2004). "Confidence estimation for machine translation", *Proceedings of the 20th International Conference on Computational Linguistics - COLING '04*, 315-es.
- Bowker, L. (2020). "Chinese speakers' use of machine translation as an aid for scholarly writing in English: A review of the literature and a report on a pilot workshop on machine translation literacy", *Asia Pacific Translation and Intercultural Studies*, 7, 1–11. Accessed on: November 29, 2023. Available at: <https://doi.org/10.1080/23306343.2020.1805843>
- Bowker, L. (2020). "Machine translation literacy instruction for international business students", *Journal of Business & Finance Librarianship*. Accessed on: January 30, 2024. Available at : https://ruor.uottawa.ca/bitstream/10393/42591/1/Bowker%20-%20WBFL1794739_AU%20-%20corrected%20proofs%20-%20July%2028.pdf
- Bowker, L. (2021). "Machine translation literacy instruction for non-translators: A comparison of five delivery formats" In R. Mitkov, V. Sosoni, J. C. Giguère, E. Murgolo, & E. Deysel (eds.) *Proceedings of the Translation and Interpreting Technology Online Conference* (p. 25–36). INCOMA Ltd. Accessed on: November 29, 2023. Available at: <https://aclanthology.org/2021.triton-1.4>
- Chrystello, C. (2023). *O uso de inteligência artificial no meio editorial*. Accessed on: January 30, 2024. Available at : <https://blog.lusofonias.net/o-uso-de-inteligencia-artificial-no-meio-editorial/>
- Diño, G. (2018). "Korean Voice Assistant Highlights Tech's Insatiable Hunger for Language Data", *Slator: Language Industry Intelligence*. Accessed on: January 28, 2024. Available at: <https://slator.com/korean-voice-assistant-highlights-techs-insatiable-hunger-for-language-data/>

Escartín, C. P., & Moniz, H. (2019). "Ethical considerations on the use of machine translation and crowdsourcing in cascading crises" In F. M. Federici & S. O'Brien (eds.) *Translation in cascading crises*. (1.^a ed.) p. 132–151. New York: Routledge.

European Commission. (2016). *Legal study on ownership and access to data*. Accessed on: January 30, 2024. Available at: https://publications.europa.eu/resource/cellar/d0bec895-b603-11e6-9e3c-01aa75ed71a1.0001.01/DOC_1

European Commission. (2021). *Artificial intelligence Act*. Accessed on January 27, 2024. Available at: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Commission. Directorate General for Communications Networks, Content and Technology. & Osborne Clarke LLP. (2016). *Legal study on ownership and access to data: Final report*. Publications Office. Accessed on: December 7 2023. Available at: <https://data.europa.eu/doi/10.2759/299944>

European Commission. Directorate General for Translation. & Bird & Bird. (2014). *Translation and intellectual property rights: final report*. Publications Office. Accessed on: December 7, 2023. Available at: <https://data.europa.eu/doi/10.2782/72107>

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Council of Literary Translators' Association. (2023). *CEATL publishes its stance on generative AI*. Accessed on: January 30, 2024. Available at: <https://www.ceatl.eu/no-one-left-behind-no-language-left-behind-no-book-left-behind>

- European Council of Literary Translators' Association. (2023). *Statement on Artificial Intelligence*. Accessed on: December 7, 2023. Available at: <https://www.ceatl.eu/wp-content/uploads/2023/11/CEATL-AI-statement-EN.pdf>
- European Parliament. (2018). *Copyright law in the EU: salient features of copyright law across the EU member states*. Accessed in: January 28, 2024. Available at: [https://www.europarl.europa.eu/RegData/etudes/STUD/2018/625126/EPRS_STU\(2018\)625126_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2018/625126/EPRS_STU(2018)625126_EN.pdf)
- European Parliament. Directorate General for Parliamentary Research Services. (2020). *The ethics of artificial intelligence: Issues and initiatives*. Publications Office. Accessed on: December 7, 2023. Available at: <https://data.europa.eu/doi/10.2861/6644>
- Frąckiewicz, M. (2023). *ChatGPT-4 and the future of literary translation: Preserving nuance and creativity*. In *TS2 SPACE*. Accessed on: December 7 2023. Available at: <https://ts2.space/en/chatgpt-4-and-the-future-of-literary-translation-preserving-nuance-and-creativity/>
- Gerken, T. & McMahon, L. (2023) "Game of Thrones author sues ChatGPT owner OpenAI", *BBC News*. Accessed on: December 7, 2023. Available at: <https://www.bbc.com/news/technology-66866577>
- Guerberof-Arenas, A., & Toral, A. (2020). "The impact of post-editing and machine translation on creativity and reading experience", *Translation Spaces*, 9(2): 255–282. Accessed on: December 7, 2023. Available at: <https://doi.org/10.1075/ts.20035.gue>
- Guerberof-Arenas, A., & Toral, A. (2022). *Creativity in translation: Machine translation as a constraint for literary texts*. In *Translation Spaces*, 11(2), p. 184–212. Accessed on: December 7, 2023. Available at: <https://doi.org/10.1075/ts.21025.gue>
- Hermas, T. (2009). "Translation, ethics, politics" In Koskinen, K., & Pokorn, N. K. (eds.). *The Routledge handbook of translation and ethics*. New York: Routledge.

- Illia, L., Colleoni, E. & Zyglidopoulos, S. (2022). *Ethical implications of text generation in the age of artificial intelligence*. Accessed on: January 30, 2024. Available at: <https://onlinelibrary.wiley.com/doi/full/10.1111/beer.12479>
- Jiang, Y. (2022). *How are neural machine-translated Chinese-to-English short stories constructed and cohered? An exploratory study based on theme-rheme structure*—ScienceDirect. Accessed on: November 22, 2023. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0024384122000791>
- Joscelyn, A. (2021). *How does AI ethics impact translation? In TAUS: The language data network*. Accessed on: January 30, 2024. Available at: <https://www.taus.net/resources/blog/multilingual-morals-how-does-ai-ethics-impact-translation>
- Kamocki, P., & O'Regan, J. (2016). *Privacy Issues in Online Machine Translation Services – European Perspective*. p. 4458 - 4462. Accessed on: December 7 2023. Available at: <https://aclanthology.org/L16-1706.pdf>
- Kenny, D. (2002). *The Ethics of Machine Translation*. In XI NZSTI National Conference. p. 1 - 10 Accessed on: January 30 2024. Available at: https://doras.dcu.ie/17606/1/The_Ethics_of_Machine_Translation_pre-final_version.pdf
- Kenny, D. (2011). *The ethics of machine translation*. Accessed on: January 30 2024. Available at: https://doras.dcu.ie/17606/1/The_Ethics_of_Machine_Translation_pre-final_version.pdf
- Kenny, D. (2022). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Accessed on: December 7 2023. Available at: <https://doi.org/10.5281/ZENODO.6653406>
- Kenny, D. & Hansen, D. (2023). *Neural Machine Translation, Large Language Models and Literary Translation: The Story So Far*. Accessed on: December 7 2023. Available at:

https://orbi.uliege.be/bitstream/2268/302554/1/presentation_ITIA_kenny_hansen.pdf

Kenny, D., & Winters, M. (2020). "Machine translation, ethics and the literary translator's voice", *Translation Spaces*, 9(1): 123–149. Accessed on: December 7 2023. Available at: <https://doi.org/10.1075/ts.00024.ken>

Loock, R., & Léchaugnette, S. (2021). "Machine translation literacy and undergraduate students in applied languages: Report on an exploratory study", *Tradumàtica: Technologies de La Traducció*, 19, 204–225. Accessed on: January 27 2024. Available at: <https://doi.org/10.5565/rev/tradumatica.281>

Moorkens, J., & Lewis, D. (2019). Chapter 28. *Copyright and the Reuse of Translation as Data*. Accessed on: January 27 2024. Available at: https://doras.dcu.ie/23974/1/Moorkens-Lewis_Chap28_Copyright-Machine-Learning.pdf

Moorkens, J., & Rocchi, M. (2020). "Ethics in the translation industry" In K. Koskinen & N. K. Pokorn (eds.) *The Routledge Handbook of Translation and Ethics* (1.^a ed.) New York: Routledge.

Munday, J. (2009). *The Routledge Companion to Translation Studies*, New York: Routledge.

Novelli, C., Taddeo, M., & Floridi, L. (2023). "Accountability in artificial intelligence: What it is and how it works", *AI & SOCIETY*. Accessed on: January 27 2024. Available at: <https://doi.org/10.1007/s00146-023-01635-y>

Nunes Vieira, L., O'Sullivan, C., Zhang, X., & O'Hagan, M. (2023). "Privacy and everyday users of machine translation", *Translation Spaces*, 12(1): 21–44.

Plouhinec, V. L. (2023). "No-one left behind, no language left behind, no book left behind", *CEATL*. Accessed on: January 27 2024. Available

- at: <https://www.ceatl.eu/no-one-left-behind-no-language-left-behind-no-book-left-behind>
- Rich, K. L. (2020). "Introduction to Ethics", *Nursing Ethics: Across the curriculum and into practice*. Accessed on: January 30 2024. Available at: https://samples.jbpub.com/9781449649005/22183_ch01_pass3.pdf
- Shen, X. (2020). *Chinese publishers see AI translation as the killer app for pushing web novels overseas*. In South China Morning Post. Access on: December 3 2023. Available at: <https://www.scmp.com/abacus/tech/article/3102891/chinese-publishers-see-ai-translation-killer-app-pushing-web-novels>
- Susam-Saraeva, Ş., & Spišiaková, E. (2021). *The Routledge Handbook of Translation and Health*. New York: Routledge.
- Taivalkoski-Shilov, K. (2019). "Ethical issues regarding machine(-assisted) translation of literary texts", *Perspectives*, 27(5): 689–703.
- University of Oxford (2023). *Tackling the ethical dilemma of responsibility in Large Language Models* | University of Oxford. Accessed on: January 30 2024. Available at: <https://www.ox.ac.uk/news/2023-05-05-tackling-ethical-dilemma-responsibility-large-language-models>
- Vale, F. (2023). "Traduções por inteligência artificial (IA) chegam a Portugal sem se fazer anunciar", *Comunidade Cultura e Arte*. Accessed on December 3, 2023. Available at: <https://comunidadeculturaearte.com/traducoes-por-inteligencia-artificial-ia-chegam-a-portugal-sem-se-fazer-anunciar>
- Van Wynsberghe, A. (2021). "Sustainable AI: AI for sustainability and the sustainability of AI", *AI and Ethics*, 1(3): 213–218
- Vardaro, J., Schaeffer, M., & Hansen-Schirra, S. (2019). "Comparing the Quality of Neural Machine Translation and Professional Post-Editing", *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, p. 1–3. <https://doi.org/10.1109/QoMEX.2019.8743218>

- Wolffe, R. (2023). "Machine translation (MT) vs generative ai for translation: pros and cons", *Lokalise*. Accessed on: January 28, 2024. Available at: <https://localise.com/blog/machine-translation-vs-ai-for-translation/>
- World Intellectual Property Organization (1979). *Berne Convention for the Protection of Literary and Artistic Works*. Accessed on January 27, 2424. Available at: <https://www.wipo.int/wipolex/en/text/283693>
- Zhan, J. (2023). *George R.R. Martin among 17 authors suing over ChatGPT*. Accessed on: November 22, 2023. Available at: <https://www.vulture.com/2023/09/george-r-r-martin-authors-guild-chatgpt-lawsuit-openai.html>

Chapter 8

Artificial Intelligence Safety as an Emerging Paradigm

Eleni Angelou¹

Artificial Intelligence (AI) safety is an interdisciplinary field undergoing transformation to become its own paradigm. By adopting a Kuhnian approach to this transformation, I seek to describe the features that make AI safety pre-paradigmatic along with the conditions that are required for the field to transition to the paradigmatic stage. I argue that the nature of the *alignment problem* – the key research question of the field – is such that it calls for its own paradigm as the research necessary to solve it goes beyond the typical desiderata in machine learning work. It incorporates a series of metascientific projects as well as breakthroughs in normative and applied ethics. The faster and better these projects are executed, the more likely that the field will come to resolve its puzzles, which will benefit society at large.

Keywords: AI Safety, Alignment Problem, Machine Learning, Kuhn, Paradigms

1. Introduction

Artificial intelligence (AI) safety is an emerging field at the intersection of computer science and philosophy, concerned with a

¹ PhD student at CUNY, Graduate Center.

pressing question: how can we reliably instill human goals in AI systems? This puzzle is also known as the *alignment problem* (Bostrom, 2016) or the *problem of control* (Russell, 2019). The increasing urgency of the problem, especially in the deep learning era (Sejnowski, 2018), calls for a careful examination of the field that is devoted to finding the solution, a field that blurs the lines between scientific, engineering, and philosophical work. Because of that observation, solving alignment requires more than incremental improvements in AI techniques.

To make this concrete, consider the following question: if we could look back from a future in which the alignment problem has been solved, what characteristics of AI safety research would have most prominently catalyzed that success? I argue that a key factor would be the maturation of AI safety into its own *paradigm*. To support this claim, I examine AI safety through the lens of Thomas Kuhn's model of scientific revolutions (Kuhn, 1962). Kuhn describes how scientific fields evolve from a pre-paradigmatic state to a structured, paradigmatic discipline with shared assumptions, methodologies, and theoretical foundations. I seek to show how AI safety is in a pre-paradigmatic state, lacking a unifying framework and a consensus on best methodologies. Before that is achieved, it is unlikely to find success in resolving the alignment problem.

The chapter explores how to best apply Kuhn's model to AI safety and discusses specific metascientific projects that could contribute to its maturation. I first outline the characteristics of AI safety at its pre-paradigmatic state, especially as this relates to the nature of the alignment problem and its implications for scientific change. Then, I propose some desiderata for AI safety to qualify as a paradigm drawing from Kuhn's epistemology of science. Finally, I offer examples of metascientific questions to ask within AI safety research.

My approach encourages exploiting the conceptual toolbox of history and philosophy of science to provide a more comprehensive understanding of risks from advanced AI models; implementing Kuhn's view on the development of AI safety is crucial for constructing an epistemological account of AI-related risks.

2. AI Safety: A Field at the Pre-paradigmatic State

The transformation of an emerging scientific field into a paradigm is marked by intrinsic uncertainty and metascientific deliberation. Uncertainty arises because the field lacks the maturity required to be studied as a discipline of its own. Maturity in this context refers to unity, clarity, and concreteness in the research agendas of scientists (Hibbert, 2016). In Kuhn's terms, at the pre-paradigmatic stage, researchers have no consensus over the body of concepts, phenomena, and techniques to use when working. They also do not follow the same rules or standards for their scientific practice, simply because no such rules or standards have been established. Establishing such a consensus is challenging – “the road to a firm research consensus is extraordinarily arduous” (Kuhn 2012, 15). Metascientific deliberation is therefore necessary to account for the uniqueness of the generation of novel scientific knowledge and practice as well as the difficulties that might be encountered on the road to normal science and puzzle-solving (Kuhn 2012).

Diagnosing the difficulties of the pre-paradigmatic phase will help illustrate how Kuhn's model applies to AI safety. The first difficulty concerns narrowing down the research directions space. Without a paradigm, or at least one obvious paradigm candidate, all hypotheses, research proposals, and agendas seem equally promising and worthy of investigation. Moreover, we observe a disunity of frameworks that concerns the concepts, theories, agendas, practices, methodological tools, and other criteria for what qualifies as having high explanatory

power. There are at least three main research directions in AI safety, but it is highly controversial whether any of them would qualify as a paradigm independently. These research directions are (1) Assistance Games, motivated in (Russell, 2019) and developed in (Fickinger *et al.*, 2020), (2) Agent Foundations (Yudkowsky, 2008), and (3) Prosaic Alignment including three proposals: HCH (Christiano, 2018), ELK (Christiano, 2022), and formalizing heuristic arguments (Christiano, Neyman and Xu, 2022). This categorization is not exhaustive although it provides a sense of the alignment research landscape and the priorities of prominent researchers. Different teams and laboratories are nevertheless optimizing for different agendas and goals as there is no agreement on which agenda is most likely to succeed.

Each of these directions reflects a different set of assumptions about AI developments. More specifically, direction (1) treats an AI system as an assistant in a reinforcement learning setup whose goal is to maximize the preference of the human user in a cooperative and interactive environment (Fickinger *et al.*, 2020). In an ideal assistance game, the AI continuously learns from the user's preferences and effectively adjusts its actions to them. Moreover, Agent Foundations aims at formulating a generalized framework of agency, such that explaining and predicting the behavior of an AI does not depend on the specifics of deep neural networks (Soares and Fallenstein, 2017). In other words, direction (2) aspires to offer mathematical tools that apply to agents of all kinds, not only those that could arise out of deep neural networks. On the opposite side, Prosaic Alignment supposes that the likelihood that powerful systems arise out of the current deep learning revolution is very high. As a result, it is important to devise tools that control deep learning models, such as the different variants of the transformer architecture (Vaswani *et al.*, 2017). The schism between the research agendas in AI safety can be articulated in terms of differing hypotheses about the nature of intelligence and the

conflicting predictions about the possibilities in the future of deep learning.

A second challenge is the widespread fundamental conceptual confusion among AI safety researchers. This rather defining feature of the pre-paradigmatic stage is explicitly expressed and acknowledged by alignment researchers working under different agendas. For instance, an Agent Foundations researcher indicates that the first step of his research plan is to “sort out our fundamental confusions about agency” (Wentworth, 2021) admitting that experts in the field share a “fundamental confusion” and there is no explicit consensus on the nature of the subject matter and the best ways to approach it.

A third challenge that characterizes pre-paradigmatic AI safety is the absence of a published corpus of work that goes together with a highly organized discipline. As Kuhn observes: “in the sciences, the formation of specialized journals, the foundation of specialists’ societies, and the claim for a special place in the curriculum have usually been associated with a group’s first reception of a single paradigm” (Kuhn 2012, 19). In other words, the institutional norms and patterns are yet to be established. This includes having a rigid definition of the group of AI safety scientists, one or more AI safety peer-reviewed journals, established conferences, standardized university curricula, and professorships.²

Despite these challenges, there are promising signs that AI safety could overcome its pre-paradigmatic phase. For example, the fact that many students and researchers have shifted their academic interests towards machine learning alignment is one way to understand the

² As of 2025, all of these have started to form, typically within the institutional circles of computer science and/or philosophy. For example, Stanford University offers CS 120: Introduction to AI Safety.

growth of the field into a scientific community. According to an estimation, there were approximately 400 people worldwide working directly on AI safety issues in 2022, including those that did not do strictly scientific work, but were adjacent to the developing paradigm of research.³ Having these researchers pursue the existing directions, conduct empirical work and experimentation to falsify hypotheses, and share their findings indicates that the field is at least receiving adequate attention to eventually have its own paradigm.

Before turning to the alignment problem itself, it is valuable to address why it is so critical to consider the development of the AI safety paradigm, among the many fields that may be undergoing transformation. AI safety stands out due to the extraordinarily high stakes of solving the alignment problem before having AI systems that could get out of control (Yudkowsky, 2008; Russell, 2019). This is primarily because whether we are able to control extremely powerful machines or transformative AI systems (TAI) might be the main factor determining the long-term future of humanity (Ord, 2020).

To better appreciate the significance of a fully formed field of AI safety, we must take into account that solving the alignment problem is increasingly becoming a priority as machine learning applications evolve into a general purpose technology, i.e., “a single generic technology, recognizable as such over its whole lifetime, that initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects” (Crafts, 2021). The risks that accompany the potential arrival of a highly powerful AI fall into two general categories: existential risks and suffering risks (Bostrom, 2016). The first kind concerns all the possible scenarios of human extinction from misaligned, super-human AI. The

³ This number has likely changed and remains difficult to estimate. See more in <https://80000hours.org/problem-profiles/artificial-intelligence/>.

second category of risks relies on hypotheses about misaligned systems that could be the cause of endless suffering for humans or other sentient beings (Tomasik, 2011). Although a full analysis of these two categories of risks exceeds the scope of this chapter, they should still be carefully considered when determining under what conditions AI safety transforms into its own paradigm.

3. The Nature of the Alignment Problem

Examining the alignment problem itself and whether it is potentially different from other scientific problems yields useful observations for the metascientific study of AI safety and for aptly situating it within the discourse on risks and technological progress. One way to think about its characteristics is to consider historical parallels in the evolution of scientific fields. In particular, the transition from alchemy to modern chemistry provides a compelling analogy for AI safety's current stage of development. Alchemists, as "proto-scientists" (Vančik, 2021), were intellectually curious yet deeply confused and uncertain about their methods and the likelihood of success. They all, however, shared a common aim summarized in this threefold: to find the Stone of Knowledge (also called "The Philosophers' Stone"), to discover the medium of Eternal Youth and Health, and to discover the transmutation of metals (Linden, 2003). Their "science" had the shape of a pre-paradigmatic field that would eventually transform into the natural science of chemistry. Importantly, their agenda ceased to be grounded upon mystical investigations as the field began to mature.

As Kuhn's model predicts, those who remained attached to the pre-paradigmatic mysticism did not get to partake in the newly formed normal science; as Kuhn remarks, "the transfer of allegiance from paradigm to paradigm is a conversion experience that cannot be forced" (Kuhn 2012, 150) which suggests that a transformation takes place facilitating the change of practice from pre-scientific to scientific.

The new field was marked by both theoretical and institutional transformation that was reflected in practices of repeatable empirical experimentation and abandoning the previously prevalent mystical pseudoexplanations.

This analogy should not be taken to imply that AI alignment retains in any sense the mystical substrate of alchemy. Rather, it shares some key structural characteristics: high uncertainty combined with attempts to work at the experimental and observational level that cannot be as well-supported as in the physical sciences. Furthermore, it shares the intention to discover something that does not yet exist with the expectation that when it does, it will make the human world substantially qualitatively different than it was prior to that invention. This belief in a technological product that will change the flow of history in extraordinary and perhaps unpredictable ways is worth remarking. It also allows us to deepen the chemistry-alignment analogy: the aim of alignment work is to make highly powerful systems follow instructions by human designers. But, at the same time, AI teams seem to be in a “technology arms race” (Armstrong, Bostrom and Shulman, 2016) to be the first to build the most powerful machine that has ever existed and could have transformative impact on humanity (Gruetzmacher and Whittlestone, 2022).

It remains valuable to ask what the history of science can teach us about the formation of new scientific fields. To the extent that AI alignment is in a pre-paradigmatic state, what insights from past scientific transitions could help accelerate its progress? Continuing with the alchemy analogy, it would be useful for the progress of alignment research to be able to trace what exactly happened when alchemy became chemistry.

Different suggestions might apply. For example, the articulation of one or more equations, or the discovery and analysis of a substance

like phlogiston. Just as phlogiston theory ultimately gave way to the discovery of oxygen, alignment research may need to identify and refine its fundamental concepts to transition into a paradigm that will find its oxygen counterpart. Scientific change, however, is rarely straightforward. As a result, drawing from past paradigm shifts to understand how AI safety could come to maturation generates uncertainties that reflect the dynamic nature of scientific change.

4. Questions about Scientific Change

A common question concerning the development of AI safety is whether the transformation from a pre-paradigmatic to a paradigmatic stage could be accelerated. This question translates into whether progress can deliberately speed up, and if so, under what conditions. Throughout the history of science, faster progress has always involved the wide acceptance of theoretical insights combined with advances in quantitative tools and effective empirical work in the field. Because of the rapid development of AI, one suggestion is to use AI models to advance alignment research.⁴ Using AI models could increase research productivity in alignment even if models cannot generate novel insights or properly theorize about scientific work in a reliable way. Such AI systems are usually large language models (LLMs) that can review bibliographical material, compose summaries, and explain concepts with less jargon, among other tasks (Wei *et al.*, 2022). While these tools do not replace human scientific reasoning, they may serve as cognitive amplifiers for certain aspects of research.

In Kuhn's framework, progress in science is discontinuous. This suggests that researchers working within the established future paradigm of alignment theory will *see* the world radically differently

⁴ This is mentioned as a strategy of leading AI labs, such as OpenAI. See more <https://openai.com/index/our-approach-to-alignment-research/>.

compared to researchers working in the paradigm of “good old-fashioned AI” (Haugeland, 1989). It is worth noticing that there were a few researchers in the old paradigm thinking that highly powerful machines will pose a problem of control, notably, I. J. Good and V. Vinge. Whether they could have predicted, however, the occurrence of the deep learning revolution (Sejnowski, 2018) and what it would imply for alignment is rather debatable.

Interestingly, even from a logical empiricist point of view, it is not the accumulation of new knowledge that will bring the new paradigm (Irzik and Grünberg, 1995). Researchers will need to recognize that the anomalies and confusions of the field at the pre-paradigmatic state require different conceptual and technical tools to be resolved. However, it is generally possible to understand the problematic aspects of the field in retrospect, i.e., once the paradigm is established.

This brings us to the concept of incommensurability which is often misunderstood in Kuhn’s epistemology of science. Incommensurability invites a twofold discussion how terms are used when placed in different theoretical contexts and about changes in worldview (Kuhn 2012; 1990). A classic example is the term “mass”; while the term appears both in the paradigm of classical mechanics and relativity theory, it connotes different meanings depending on the paradigm. But while some concepts can be translated from an older paradigm to the vocabulary of the new one, it is generally impossible to find a one-to-one correspondence between the terms of the old paradigm and the terms of the new one. This semantic dependence on the theory implies that there is no common measure although a comparison remains possible (Kuhn 1990). In that sense, as AI safety continues to evolve, we may find that concepts such as agency or intelligence are incommensurable with what was used to describe systems in the old AI paradigm. This will not suggest that there is no

way to neutrally compare e.g., the concept of intelligence without anchoring it into a theoretical apparatus. It is also possible that some terms preserve their meanings despite the paradigm change (Kuhn 1990, 36).

The change of scientific worldview is extensively discussed by Kuhn (Kuhn, 1962, 1990) and by Hanson who specifically explains the notion of theory-ladenness (Hanson, 1958). Kuhn draws a parallel between the change of worldview to how we suddenly *see* different shapes in gestalt figures (Kuhn 2012, 114). This description of scientists seeing the world in different ways implies the automaticity of perceptual change: the scientist cannot control the shift of attention that has occurred. A typical example of this in the history of science is astronomy. Notably, while western Europeans started discovering planets the half century after the introduction of the Copernican paradigm, the Chinese were long aware of many more stars in the sky. This is because their cosmological paradigm offered a different model of celestial change.

Kuhn explains how he originally conceived the idea of a theory or a thinker belonging to a different paradigm. He describes sitting at his desk with Aristotle's *Physics* open in front of him when "suddenly the fragments in my head sorted themselves out in a new way, and fell into place together [...] for all at once Aristotle seemed a very good physicist indeed, but of a sort I'd never dreamed possible" (Kuhn 2014). From that point onwards, Kuhn would argue that scientists do not try to do the job of their predecessors simply in a better way than they did (Reisch, 2016); they see a different world, they attend to different phenomena and features of reality. This leads to an asymmetry in that they operate within a completely unique context and thus construct different world models and provide different explanations.

On a similar note, but with a cognitive focus, Hanson describes the phenomenon of theory-ladenness to illustrate that scientific understanding depends on perceptual input and its processing. Hanson famously quotes Duhem's example:

Enter a laboratory; approach the table crowded with an assortment of apparatus, an electric cell, silk-covered copper wire, small cups of mercury, spools, a mirror mounted on an iron bar; the experimenter is inserting into small openings the metal ends of ebony-headed pins; the iron oscillates, and the mirror attached to it throws a luminous band upon a celluloid scale; the forward-backward motion of this spot enables the physicist to observe the minute oscillations of the iron bar. But ask him what he is doing. Will he answer 'I am studying the oscillations of an iron bar which carries a mirror'? No, he will say that he is measuring the electric resistance of the spools. If you are astonished, if you ask him what his words mean, what relation they have with the phenomena he has been observing and which you have noted at the same time as he, he will answer that your question requires a long explanation and that you should take a course in electricity (Hanson 1958, 16-17).

Both from Kuhn and Hanson, it becomes clear that participating in a paradigm means to learn to attend to certain phenomena and features of the parts of the world that are under investigation. At the same time, this implies learning to block out everything else; part of belonging to a paradigm means that scientists agree on what to attend to and what to block out.

Throughout his writings, Kuhn challenges the idea of scientific progress. Characteristically, he writes: "We must explain why science — our surest example of sound knowledge — progresses as it does, and we must first find out how in fact it does progress" (Kuhn 1970,

20). It is plausible to argue that Kuhn would be skeptical of a potential technological singularity (Sandberg, 2013) or the completion of science. In Kuhn's model, science does not move towards absolute knowledge, so why would there be an end? Since the scientific process is discontinuous, it is not meaningful to argue that science aims at "objective" truth or simply truth as correspondence to reality (Rorty, 2003). Paradigms are incommensurable, therefore there is no neutral way to talk about how one paradigm finds "more truth" than the other. One might assume that Kuhn purposefully talks very little about the notion of truth in the Structure (Bird, 2012). He returns to the topic in the Postscript to the second edition mostly to address the various criticisms against the Structure about relativism (Kuhn 2012, 204). While assessing Kuhn's understanding of truth is out of the scope of this paper, it is necessary to note that Kuhn seems to be at least a neutralist stance about truth (Bird, 2007). This has in many instances made him a major ally of antirealist views in science while it is not implausible to find ways in which the Kuhnian model is compatible with scientific realism.

5. Desiderata for the New Paradigm

The fact that people can form coalitional groups and agree to conform to certain rules and norms does not make their group nor their work a scientific paradigm. It is thus essential to sketch out what we can expect AI safety to look like as a paradigm, based on Kuhn's model of the structure of science. This can be regarded as experimenting with a future history of AI safety – the question of looking back from a future of successful AI alignment with all the ambiguity it involves. As such, it is useful to lay out a series of desiderata that AI safety will qualify for once it has entered its paradigmatic stage. First, while the term *paradigm* has a long history that dates back to Aristotle's Rhetoric (Hacking, 2012), paradigms are defined in the Structure as "accepted

examples of actual scientific practice—examples which include law, theory, application, and instrumentation together—[that] provide models from which spring particular coherent traditions of scientific research” (Kuhn 2012, 11). In that sense, the paradigm will shape scientific work in its entirety, from education and recruiting new researchers to everyday methods and practices.

It follows that the function of a paradigm is both cognitive and normative (Kindi and Arabatzis 2013, 91) which means that paradigmatic AI safety will 1) prepare students for becoming professional researchers and members of a scientific community, 2) select the class of facts that “are worth determining both with more precision and in a larger variety of situations” a process characteristic of normal science also called “fact gathering” (Kuhn 2012, 25-27), 3) define the specific problems that must be solved (*Ibid*, 27-28), 4) offer criteria for selecting these problems (*Ibid*, 37), 5) guarantee the existence of stable solutions to these problems (*Ibid*, 28), 6) provide methods and standards of solutions, 7) aim at quantitative laws, 8) make predictions, and 9) give satisfactory explanations for phenomena.

It is worth remarking that the paradigm offers what Kuhn calls a “theoretical commitment” without which important facets of scientific activity do not exist, namely the discovery of laws. Kuhn’s examples highlight that the examination of measurements alone is never enough for finding quantitative laws. The history of science provides sufficient evidence in favor of the claim that the experimental, Baconian method does apply, but only to a certain degree. Kuhn mentions Boyle’s Law, Coulomb’s Law, and Joule’s formula (*Ibid*, 28) to argue that they were discovered through the application of a particular scientific paradigm – even if it was not as explicit at the time – and not through the study of measurements. In that sense, the

theoretical commitment encapsulates a set of assumptions and concepts which are a prerequisite for the discovery of laws. By analogy, we can anticipate AI safety to find and formulate laws or scientific models once the theoretical commitment is clearly established and broadly accepted.

6. Metascientific Projects

6.1. Conceptual clarification

The first question in metascientific inquiry is whether and how can philosophy contribute to AI safety in a concrete and practical way. While the role of philosophy in scientific progress is a matter of debate, it is arguable that philosophy can, to some extent, assist in disambiguating central concepts and frameworks that shape the field. AI safety, in particular, presents many conceptual difficulties that stem from the intrinsically perplexing nature of the study of intelligence and agency. For example, the research agenda of Agent Foundations is focused on modeling agents so that future powerful models exhibit behaviors we can predict and consequently, control. Philosophers working on the alignment problem can contribute to the conceptual clarification and modeling of agency using analytic tools that are common in philosophical reasoning.

There is one important objection to the claim in favor of the contribution of philosophical conceptual clarification. Looking at the history of science more generally, it does not seem like philosophy typically helped directly with scientific research. In other words, while some work in the philosophy of science engages with specific problems in the sciences, it is rare that it actually accelerates the progress of the field by disambiguating fundamental concepts beyond merely semantic disagreements. For that to be a true possibility,

philosophers would have to work as scientists and scientists would have to cultivate a philosophical disposition.

A response to this objection is that throughout the history of science and across various disciplines, scientists have in many instances acknowledged the necessary role of philosophical thinking in articulating the right questions and trying to explain the world rationally. Concepts are often understood as the “building blocks” of theories (Gerring, 1999). In that sense, as Popper notes, the method of philosophy and science is a single one, and that is rational discussion (Popper 2002, xix). More specifically on the relationship of AI and philosophy, McCarthy hints at the need for conceptual clarification aided by working on philosophical questions (McCarthy, 2008). This is especially useful since both fields are concerned with concepts such as goals, knowledge, belief, etc. Inquiries into AI safety are different from the typical work of a machine learning engineer precisely because this is a normative inquiry about how to ensure that these new cognitive systems will be beneficial for humanity. At the same time, the creation and deployment of highly advanced AI raises a series of questions about moral patienthood through the potential of digital consciousness (Shiller, 2024). As such, AI safety incorporates ethics in both the normative and applied dimension.

A central problem at the basis of AI concerns the conditions that will allow testing intelligent behavior for example, by applying the Turing test. Analyzing what the Turing test is about requires philosophical rigor so as to determine and interpret the criteria to pass the test. For that reason, it has been a common theme in philosophy e.g., in (Searle, 1980; Dennett, 1984; Chalmers, 1994). It overall appears that conceptual clarification is indispensable to rational discourse and the more thoroughly that it is executed, the more likely it is to yield itself conducive to paradigm transformation.

6.2. Questions about the nature of machine learning

Another project in the philosophy of science adjacent to the development of the AI safety paradigm focuses on questions about the metaphysics of machine learning and specifically, the nature of deep neural networks. This discussion is centered around whether we can emulate the functions of human minds by imitating the neural circuitry of the human brain. Until recently, this seemed unlikely to work. Steven Pinker criticized connectionism in the early 2000s arguing that

Humans don't just loosely associate things that resemble each other, or things that tend to occur together. They have combinatorial minds that entertain propositions about what is true of what, and about who did that to whom, when and where and why. And that requires a computational architecture that is more sophisticated than the uniform tangle of neurons used in generic connectionist networks. (Pinker, 2003, 79)

Pinker goes on to expose the limitations of connectionism that derive from this not sufficiently sophisticated computational architecture of neural networks. All this is to show that a complete human thought cannot be represented in a generic network that is used in machine learning. Specifically, neural networks would not be able to distinguish between kinds and individuals, have thoughts that are not merely a summation of smaller parts (compositionality), play with logical quantification, embed one thought in another (recursion), and reason categorically (Pinker, 2003).

The most recent development of machine learning proved all these claims false. Deep neural networks are surprisingly good at all the tasks Pinker mentions. In particular, large language models or transformers (Vaswani *et al.*, 2017) and other deep learning

architectures have generated impressive results such as writing like humans or making scientific discoveries, e.g., AlphaFold (Jumper *et al.*, 2021). While there is no consensus on why these models are so successful (Ngo, 2022), it is plausible to conclude that intelligence was not so hard to find, after all. How they work seems to be itself a problem between philosophy, science, and engineering. One or more solutions to it will likely be at the foundation of the new AI safety paradigm. Explainable AI, i.e., models that are not regarded as black boxes but are rather *interpretable* (Li *et al.*, 2022) would overall make safety easier to achieve.

Lastly, considerations about whether it is possible for machine learning systems to think or become highly competent at human-level tasks go back to AI researchers in the 1950s (McCarthy, 2008) and remain relevant even in AI textbooks such as in (Russell and Norvig, 2009). Gaining clarity on how neural networks work and what accounts exactly for their success is almost equivalent to unlocking the nature and conditions of cognition.

7. Conclusion

The emergence of AI safety as a paradigm in the making fuels the study of how new science is generated and presents an opportunity to zoom in on some of the most fundamental questions about the nature of scientific practice and the epistemology of progress in science and technology. My aim was to explore the application of Kuhn's theory of the structure of science in the case of AI safety. I first argued that the field is at its pre-paradigmatic state which motivates a series of metascientific observations. For these observations to make sense, I offered an account of the nature of the alignment problem, the central problem AI safety must tackle. In doing so, I suggested that there is an analogy between the development of AI safety and the history of chemistry. I then sketched out the requirements for the field to

transform into a fully evolved paradigm according to Kuhn's view on scientific change.

As AI safety is undergoing quick development, it is reasonable to speculate that it may transition to the paradigmatic phase in the near-term future, although there remains a lot of uncertainty as to what breakthroughs would be necessary for solving alignment. A potential way to accelerate this transformation is by investing in philosophically-minded, metascientific projects such as conceptual clarification and the theoretical study of machine learning. If targeted at the right questions, they could be conducive to transitioning to a paradigm of AI safety. The need to reach the paradigmatic state becomes more and more urgent as machine learning capabilities advance, and models exhibit intelligent behavior that could eventually pose serious risks for society at large.

References

- Armstrong, S., Bostrom, N. and Shulman, C. (2016) "Racing to the precipice: a model of artificial intelligence development", *AI & Society*, 31(2): 201–206.
- Bird, A. (2007) "What is scientific progress?2, *Noûs*, 41(1): 64–89.
- Bird, A. (2012) "10 Kuhn, Naturalism, and the Social Study of Science" in V. Kintē and T. Arabatzis (eds.) *Kuhn's the Structure of Scientific Revolutions Revisited*. New York: Routledge.
- Bostrom, N. (2016) *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford, United Kingdom; New York: Oxford University Press.
- Chalmers, D.J. (1994) "On implementing a computation", *Minds and Machines*, 4(4): 391–402.

- Christiano, P. (2018) *Humans consulting HCH, Medium*. Available at: <https://ai-alignment.com/humans-consulting-hch-f893f6051455> (Accessed: 12 December 2022).
- Christiano, P. (2022) *Eliciting latent knowledge, Medium*. Available at: <https://ai-alignment.com/eliciting-latent-knowledge-f977478608fc> (Accessed: 12 December 2022).
- Christiano, P., Neyman, E. and Xu, M. (2022) "Formalizing the presumption of independence", *arXiv preprint arXiv:2211.06738* [Preprint].
- Crafts, N. (2021) "Artificial intelligence as a general-purpose technology: an historical perspective", *Oxford Review of Economic Policy*, 37(3): 521–536.
- Dennett, D.C. (1984) "Can machines think?" In M.G. Shafto (ed.) *How We Know*. New York: Harper & Row.
- Fickinger, A. et al. (2020) "Multi-principal assistance games", *arXiv preprint arXiv:2007.09540* [Preprint].
- Gerring, J. (1999) "What makes a concept good? A criterial framework for understanding concept formation in the social sciences", *Polity*, 31(3): 357–393.
- Gruetzemacher, R. and Whittlestone, J. (2022) 'The transformative potential of artificial intelligence', *Futures*, 135, 102884.
- Hacking, I. (2012) 'Introductory essay', N/D.
- Hanson, N.R. (1958) *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. 1st edition. Cambridge: Cambridge University Press.
- Haugeland, J. (1989) *Artificial Intelligence: The Very Idea*. Reprint edition. Cambridge, Mass: Bradford Books.
- Hibbert, R. (2016) "What Is an Immature Science?", *International Studies in the Philosophy of Science*, 30(1), pp. 1–17.

- Irzik, G. and Grünberg, T. (1995) "Carnap and Kuhn: Arch enemies or close allies?", *The British Journal for the Philosophy of Science*, 46(3): 285–307.
- Jumper, J. et al. (2021) "Highly accurate protein structure prediction with AlphaFold", *Nature*, 596(7873): 83–589.
- Kindi, V. and Arabatzis, T. (2013) *Kuhn's The structure of scientific revolutions revisited*. New York: Routledge.
- Kuhn, T.S. (1962) *The Structure of Scientific Revolutions: 50th Anniversary Edition*. 4th edition. Chicago; London: University of Chicago Press.
- Kuhn, T.S. (1970) "Logic of discovery or psychology", *Criticism and the Growth of Knowledge*, 4: 1.
- Kuhn, T.S. (1990) "The road since structure", *PSA: Proceedings of the biennial meeting of the philosophy of science association*, Philosophy of Science Association, pp. 3–13.
- Li, Xuhong et al. (2022) "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond", *Knowledge and Information Systems*, pp. 1–38.
- Linden, S.J. (2003) *The Alchemy Reader: From Hermes Trismegistus to Isaac Newton*. Cambridge University Press.
- McCarthy, J. (2008) 'The Philosophy of AI and the AI of Philosophy', In *Philosophy of Information*. Elsevier, pp. 711–740.
- Ngo, R. (2022) "The alignment problem from a deep learning perspective", *arXiv preprint arXiv:2209.00626* [Preprint].
- Ord, T. (2020) *The Precipice: Existential risk and the future of humanity*. New York: Hachette Books.
- Pinker, S. (2003) *The Blank Slate: The Modern Denial of Human Nature*. New York: Penguin.
- Popper, K. (2002) *The Logic of Scientific Discovery*. 2nd edition. London: Routledge.

- Reisch, G.A. (2016) "Aristotle in the cold war: On the origins of Thomas Kuhn's the structure of scientific revolutions", *Kuhn's structure of scientific revolutions at fifty: Reflections on a science classic*, pp. 12–30.
- Rorty, R. (2003) "Dismantling truth: Solidarity versus objectivity", *The Theory of Knowledge: Classical and Contemporary Readings*, [Preprint].
- Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin Books.
- Russell, S. and Norvig, P. (2009) *Artificial Intelligence: A Modern Approach*. 3rd edition. Upper Saddle River: Pearson.
- Sandberg, A. (2013) "An overview of models of technological singularity", *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*, pp. 376–394.
- Searle, J. (1980) "Minds, brains, and programs", *Behavioral and Brain Sciences*, 3(3): 417–57.
- Sejnowski, T.J. (2018) *The Deep Learning Revolution*. Illustrated edition. Cambridge, Massachusetts: The MIT Press.
- Shiller, D. (2024) "Functionalism, integrity, and digital consciousness", *Synthese*, 203(2): 47.
- Soares, N. and Fallenstein, B. (2017) "Agent foundations for aligning machine intelligence with human interests: a technical research agenda", *The technological singularity: Managing the journey*, pp. 103–125.
- Thomas Kuhn (2014) "What Are Scientific Revolutions?" In *Philosophy, Science, and History*. New York: Routledge.
- Tomasik, B. (2011) "Risks of astronomical future suffering", *Foundational Research Institute: Berlin, Germany* [Preprint].
- Vančik, H. (2021) "Alchemy" In *Philosophy of Chemistry*. Cham: Springer.

- Vaswani, A. *et al.* (2017) "Attention is all you need", *Advances in neural information processing systems*, 30.
- Wei, J. *et al.* (2022) "Emergent abilities of large language models", *arXiv preprint arXiv:2206.07682* [Preprint].
- Wentworth, J. (2021) "The Plan", *LessWrong* [Preprint]. Available at: <https://www.lesswrong.com/posts/3L46WGauGpr7nYubu/the-plan>.
- Yudkowsky, E. (2008) "Artificial Intelligence as a Positive and Negative Factor in Global Risk" In *Global Catastrophic Risks*. Oxford: Oxford University Press.

Chapter 9

Managing the Inductive Risk of Ethical Decision-making with Autonomous Systems: A Problem of Cost-Sensitive Optimization

Koray Karaca¹ and Pablo Muñoz²

Autonomous systems are a topic relevant to both philosophy of science and artificial intelligence (AI) ethics. On the side of philosophy of science, grounded on the concept of inductive risk, researchers have been interested in value-driven scientific decision-making processes performed in the machine learning (ML) modelling tasks part of the design, implementation and use of these systems (Karaca, 2021; Biddle, 2020). On the side of AI ethics, researchers and practitioners, under the framework of Responsible Research and Innovation (RRI), have focused on how to make autonomous systems socially acceptable or publicly trustworthy. We propose a new approach for achieving public trustworthiness in autonomous systems based on performing public cost-sensitive optimization in the development of these systems. In the first place, we show that the two approaches proposed and applied until now, the participatory and the transparent

¹ Koray Karaca is an assistant professor in the Department of Philosophy at the University of Twente in the Netherlands.

² Pablo Muñoz is a Ph.D. candidate in the Department of Philosophy at University of the Andes in Colombia and works as Innovation Management Professional at Compensar.

approach, recognize that public trustworthiness in autonomous systems is associated with the management of inductive risk arising in these systems. However, these approaches do not address adequately this management. In the second place, we highlight that autonomous systems function through ML modelling, and, because of this, inductive risk must be managed through cost-sensitive optimization. In this way, we demonstrate that the framework proposed by the scholars of value-driven scientific decision-making for managing ML inductive risk, i.e., the cost-sensitive optimization framework, can be applied to achieve public trustworthiness in autonomous systems applications. Consequently, we emphasize the importance of employing research done in philosophy of science for complementing AI ethics and RRI.

Keywords: inductive risk, machine learning, cost-sensitive optimization, autonomous systems, public trustworthiness.

1. Introduction

Autonomous systems are a topic relevant to both *philosophy of science* and *artificial intelligence (AI) ethics*. On the side of philosophy of science, researchers have been interested in the *value-driven scientific decision-making* processes performed in the *machine learning* (ML) modelling tasks part of the design, implementation and use of these systems (Karaca, 2021; Biddle, 2020). On the side of AI ethics, researchers and practitioners, under the framework of *Responsible Research and Innovation* (RRI), have focused on how to make autonomous systems *socially acceptable* or publicly trustworthy (Jedličková, 2024; Leikas, Koivisto and Kotcheva, 2019). Scholars interested in the value-driven scientific decision-making performed in ML modelling have established that a way to ensure that this decision-making is guided by human or social values is to perform *cost-sensitive optimization* processes in these modelling activities (Karaca, 2021; Elkan, 2001). At

the same time, AI ethicists have acknowledged that the social acceptability of autonomous systems can be achieved solely through *enhanced public trust*. Also, these ethicists have pointed out that attaining enhanced public trust is related to the *management of inductive risk* (Carrier and Izirk, 2021; Irzik and Kurtulmus, 2021). As value-driven scientific decision-making occurring at ML modelling is also an inductive risk management problem, philosophical research on value-driven decision-making has been employed to achieve this type of trust in autonomous systems projects. However, the RRI approaches used by AI ethicists for applying this research has not considered cost-sensitive optimization as a central aspect of this application.

In this chapter, we argue that not articulating this concept of cost-sensitive optimization has limited these RRI approaches in their task of assuring public enhanced trust in autonomous systems' projects. Thus, 1) we demonstrate that cost-sensitive optimization can be employed for achieving this trust. 2) We reveal that it is not enough to include public stakeholders in scientific decision-making tasks or to use transparent rule-based reasoning in this decision-making. We state that it is crucial to involve these stakeholders in cost-sensitive optimization tasks for achieving enhanced public trust and social acceptability in these systems' projects. 3) We demonstrate that environmental data processed by autonomous systems and unexpected situations occurring in these systems are not an obstacle to the employment of cost-sensitive optimization in these systems. For achieving these purposes, we, first, relate, from a philosophy of science perspective, inductive risk in ML and cost-sensitive optimization (section 2). Also, we justify why inductive risk must be managed in autonomous systems. Second, we establish why the social acceptability of autonomous systems is a problem of enhanced public trust and, thereby, of inductive risk management (section 3). Fourth,

we expose the two main RRI approaches for addressing enhanced public trust, the participatory and the transparent approach, and highlight their limitations (section 4). Fifth, we propose a third approach (section 5). Sixth, we address some objections to our approach (section 6).

2. Inductive Risk in ML

2.1 Inductive risk and cost-sensitive optimization

Inductive risk is a concept employed by philosophers of science to address the relationship between the acceptance or rejection of scientific hypothesis, the social impact of this acceptance/rejection and the evidence available that supports this acceptance/rejection (Karaca, 2021; Douglas, 2000). Philosophers of science highlight that, in scientific activities, the evidence that supports a decision of acceptance or rejection is always limited. Therefore, scientists can make a false acceptance (false positive) or a false rejection (false negative). Additionally, philosophers of science are aware that some scientific activities are embedded in social contexts, and, because of this, their acceptance/rejection decisions and the errors of falsely accepting/rejecting might have social impacts.

Philosophers of inductive risk show that in this social practice of science, while some results are more favorable than others because they bring social benefits, other results are less favorable or harmful because they produce negative social impacts (Karaca, 2021; Douglas, 2000). In this way, scientists classify these acceptance/rejection results according to the *social or human values* that determine which of these results is favorable and which is harmful. For example, toxicologists working on a plastic-regulatory agency seek to preserve and promote safety among plastic users and consumers. Thus, when accepting or rejecting whether a substance for manufacturing plastic products is

toxic, they are conscious that accepting that the substance is toxic will make plastic manufacturers look for alternative substances or increase the accuracy of their toxicity testing procedures. In this way, this acceptance will promote the value of safety among the plastic industry and will ensure that safe plastics are used and consumed.

In this case, a true acceptance is beneficial, and a true rejection is not beneficial (a true rejection would allow manufacturers to use this substance and not look for substitutes whose toxicity is not disputed). Also, a false rejection could be very harmful, since a toxic substance would be used for manufacturing plastic products. Instead, a false acceptance would not be harmful and could be beneficial, since it would increase the accuracy of toxicity tests or promote the search for alternative substances. This characterization of socially embedded scientific activities shows that scientists must determine whether the available evidence is strong enough to make an acceptance/rejection decision based on their social values. Following the previous example, if 60% of the testing samples show that the substance is not toxic, toxicologists can state that this evidence is not strong enough and that they require that 90% of the testing samples show that the substance is not toxic to reject toxicity. In this way, based on the value of safety, scientists determine how to manage limited evidence for accepting/rejecting toxicity in the mentioned substance and, consequently, reduce the probability of false negatives. Therefore, we state that, in socially embedded science, social values *guide* or *drive* scientists in the management of the available evidence for acceptance/rejection decisions and the inductive risks attached to this decision-making.

Inductive risk pervades not just scientific hypothesis acceptance/rejection, but also *scientific mathematical models* acceptance/rejection (Karaca, 2021; Winsberg, 2012). In mathematical modelling, scientists collect and process a group of data to evaluate which input-output

function this group of data reveals, i.e., given certain input, which output it produces. Inductive risk might arise in scientific mathematical models because the data used for finding an input-output function is always limited. In this sense, a group of data is used for supporting the acceptance/rejection of an input-output function or modelling solution. Following the earlier example, scientists might evaluate whether the input 'high concentrations of molecule X' produce the output 'high toxicity in the plastic substance A' based on the experimental data gathered in a toxicological test. For this reason, the management of limited evidence and of inductive risk in scientific mathematical modelling done in a social context is guided by social values.

ML modelling is a type of scientific mathematical modelling (Karaca, 2021). ML modelling is a process which consists of two main steps. 1) In the training step, a learning algorithm is executed by a computer program. This algorithm is the set of instructions established by the modelers for analyzing a data set based on a combination of specific parameters and features. Therefore, this algorithm finds the input-output function that, based on the parameters and features indicated by the modelers, approximates the most to the unknown input-output function that determines exactly, given a specific input, what the output is. 2) In the evaluation step, modelers iterate the same algorithm in a second dataset and evaluate whether the modelling function found by the algorithm has the same approximation rate as the function found at the training step. When the approximation rates are different, modelers change the parameters upon which the algorithm finds the function until they produce a similar approximation rate. This process of adjusting the parameters is known as ML optimization. The optimization of the learning algorithm allows scientists to generalize the modelling solution found by the algorithm. In fact, the minimization of the differences between

the approximation rates of the solutions found by the algorithm at the training and testing stages indicates that the features-parameters combination, when applied in multiple datasets, produce the same modelling solution.

In ML applications that might have societal impacts (e.g. credit scoring systems and oncological diagnostics), inductive risk is managed through *cost-sensitive optimization* (Karaca, 2021). In the same way as other types of mathematical modelling applied in activities that might produce social impacts, scientists, based on social values, must determine whether the empirical evidence is strong enough to accept/reject a modelling solution. Therefore, they attach a cost to each of the acceptance/rejection decisions and the inductive risks arising in them, and they manage evidence aiming at choosing the modelling solution with less costs. As the management of evidence consists in selecting the features and parameters upon which the algorithm finds a modelling solution, scientists choose the features-parameters combination based on the costs of the modelling solution produced by each combination. Thus, at the training stage, the algorithm will analyze a dataset and find a modelling solution; at the evaluation stage, the algorithm will iterate again the analysis in a second data set and find a second solution. Subsequently, modelers will optimize the approximation rates of the two functions not just by seeking a features-parameters combination that minimizes the difference between these rates. Instead, they will select this combination based on two conditions: 1) the combination minimizes the rates; 2) the combination produces the modelling solution with less costs. In other words, the optimization process is sensitive to the costs attached to each of the results of the model acceptance/rejection decision-making.

2.2 Inductive risk in autonomous systems

The most important feature of autonomous systems is their *situational awareness* (Fernández et al., 2021). These systems are arrays of sensors and actuators. Thus, through their sensors they capture environmental data and adapt their configuration according to these captured data. This process of adjusting the characteristics of the system to the surrounding environment gives autonomous systems their main feature of being situationally aware. For example, an autonomous vehicle is equipped with cameras that enable it to detect objects near it. When it detects the object, it sends a signal to an actuator connected to the braking system. Thus, the vehicle automatically brakes.

Environmental data needs to be processed in real time to enable autonomous systems to change their configuration according to the surroundings environment (Su et al., 2024; Sharma, 2021). This data processing task cannot be done manually. Autonomous systems designers and implementers need *ML solutions* to enable these systems to capture environmental data, process it, and adapt their configuration. Therefore, autonomous systems are imbued with ML models and consequently, they are vulnerable to inductive risk. We use the case of autonomous vehicles to show how this risk arises in these systems: the vehicle is driving in a highway and detects a rock straight ahead. The vehicle brakes. The reason why the vehicle breaks is that the ML model performing the environmental data processing has been trained and evaluated by a series of images of rocks. Therefore, it has found a certain generalizable modelling function that states that, each time the vehicle detects a rock with certain characteristics and located at certain distance from the vehicle, the output is a brake command. This modelling solution might produce false positives or false negatives. The vehicle might brake when there is no rock ahead. Also, it might not brake even when there is a rock

ahead. Therefore, as the process of adapting to the environment of autonomous systems is supported by ML models, inductive risk might arise in these systems.

It is essential to indicate that autonomous systems are socially embedded ML applications (Su et al., 2024; Fernández et al., 2021). One example of socially embedded autonomous systems is autonomous driving. The use of these vehicles in multiple transportation applications, such as freight delivery and urban mobility, requires assuring the *safety* of the multiple stakeholders involved in this use (e.g., passengers and pedestrians). A false negative regarding the detection of a pedestrian might lead to a collision between the vehicle and the pedestrian. Therefore, in these applications human values guide the model acceptance/rejection decisions and these decision-making processes are managed through cost-sensitive optimization. In the case of autonomous driving, model acceptance/rejection should be managed by selecting the parameters-features combination that has less safety costs.

3. Inductive Risk and Public Enhanced Trust

Philosophy of inductive risk is not the only philosophical field that can be employed for assessing the social impacts of autonomous systems. AI ethics has as its main purpose ensuring that the distinct decision-making aspects related to the *development of AI technological applications* (design, implementation and use), among them autonomous systems, is guided by human or social values (Jedličková, 2024; Leikas, Koivisto and Kotcheva, 2019). Thus, AI ethics is a design methodology (i.e., a methodology for guiding technological development) that aims to regulate the application of autonomous systems based on social values.

One aspect of autonomous systems applications highlighted by AI ethics is that these systems can be guided by social values only if they have *social acceptability* (Leikas, Koivisto and Kotcheva, 2019, p. 18). Among the sub-methodologies of AI ethics, Responsible Research and Innovation (RRI) is the one that addresses the requirement of social acceptability (Carrier and Izirk, 2021).³ The experts in the application of RRI point out that, to effectively ground scientific and technological applications in social values, it is necessary to (1) involve the public in the scientific and technological projects and practices. (2) It is necessary to ensure that the ends of these scientific and technological applications are ends pursued by society. Carrier and Izirk (2021) indicate that the first requirement is known as the requirement of “science with society”. The second requirement is known as “science for society”.

RRI practitioners claim that accomplishing the first requirement is related to *public enhanced trust* (Irzik and Kurtulmus, 2021). RRI practitioners state that scientific and technological activities should be social activities where there is *interaction* between two main groups. The first group is the group of *experts* or scientists. The main characteristic of this group is that it *produces and communicates information* to the second group, because the members of the second group cannot produce this information by themselves. Also, this information is crucial for the preservation of social values and for the managing of social impacts. The second group is the group of social or *public agents*. This group defends certain social values and is vulnerable to certain social impacts. For example, in autonomous

³ *Responsible research and innovation* is fundamentally a concept that emphasizes the importance of giving access and enable participation in science and innovation to the multiple actors of society. Therefore, this concept has given place to multiple methodologies for improving the relationship between scientific researchers, innovators and the public (see Carrier and Izirk, 2021).

vehicles, users (second group) cannot know by themselves the weather conditions of the road. They depend on climate and technological experts (first group) that measure these conditions and send the data to the memory of the autonomous vehicles via wireless communication. As these users' lives depend on the safety driving of autonomous vehicles, they defend the value of safety and are vulnerable to safety breaches.

Also, the information about weather gathered by experts is crucial for assuring safety. If the road is very foggy, it is preferable not to use the autonomous vehicle. Thus, RRI practitioners characterize scientific and technological activities as activities where information circulates from one group to another, because this second group cannot obtain this information by itself, but this information is, in any case, necessary for the wellbeing of the second group. For this reason, RRI practitioners consider that scientific activities are a locus of *public trust*. In the first place, public agents depend on the other group for the defense of their values and for the preservation of their wellbeing. In the second place, public agents believe that the experts have goodwill and, consequently, will communicate the information needed for this defense and this preservation. Both conditions set the two groups in a relationship of trust, where the public agents group trusts the other group, i.e., depends on the other group and believes in its goodwill.

However, RRI practitioners argue that the trust the public has in the experts is not *basic trust*. By contrast, this trust is *enhanced* (Irzik and Kurtulmus, 2021). RRI practitioners recognize that scientific and technological activities are activities where inductive risk must be managed. In fact, the information produced and communicated by experts follows the form of scientific products and methodologies, such as hypotheses and models. Therefore, this information consists

in hypotheses or models that are accepted or rejected based on a body of limited evidence. Consequently, the production and communication of information done by the experts is vulnerable to false positives and negatives. Following the previous example, climate experts might claim that the fog will dissipate within an hour based on the available climate data and err in this prediction.

The fact that inductive risk might arise in the trust relationship between scientists and the public reveals that an additional component needs to be attributed to this trust relationship (Irzik and Kurtulmus, 2021). In the basic trust relationship, the public depends on the scientists for the public's wellbeing and beliefs that scientists will act in a way that will protect this wellbeing. However, as scientists might err in their scientific activities due to limited evidence, they have an interest in avoiding errors. Therefore, this interest is different from the interest of preserving the public's wellbeing. Also, this interest might conflict with this second interest. For instance, to avoid making a false claim that the fog will not dissipate on the road, climate experts might claim that 60% of probability that the fog will dissipate is enough for claiming that the road will not be foggy in the coming hours and that there are favorable conditions for autonomous driving.

This decision might impact negatively on users' wellbeing, since it might lead users to ride autonomous vehicles in foggy and dangerous roads. For this reason, RRI practitioners recognize that inductive risk can be managed according to social and ethical values. They point out that scientists can establish the threshold that determines the strongness of evidence based on social values. Thus, the interest of scientists in avoiding errors can *match* the interest of the public. In the case of foggy roads and autonomous driving, scientists can favor safety (an interest of the public) and, consequently, situate the threshold of evidence for accepting that fog will disappear in a high

position (e.g., around 90% of probability). This matching is the component that turns public trust relationships into enhanced trust relationships. In basic trust relationships, public agents depend on the experts for their wellbeing and believe that they will protect this wellbeing. As science is pervaded with inductive risk, the interests of scientists or experts might conflict with the interest of the public. Nonetheless, scientists might match their interest with the public's interest by means of the management of inductive risk based on social values. Hence, in enhanced trust relationships, public agents also believe that scientists will make this match in their scientific activities.

4. The Two Approaches for Enhanced Public Trust in Autonomous Systems Projects

The participatory approach

To achieve public enhanced trust in autonomous systems projects, RRI practitioners have proposed and applied multiple methods (Tekkesinoglu et al., 2025; Patel et al., 2024; Leonard and Tochia, 2022; Roberson et al., 2022). We consider that these methods can be classified into two main approaches to achieve this type of trust in socially embedded autonomous systems. The first approach is the participatory approach. RRI practitioners advocates of this approach show that in projects related to the development of autonomous systems the trust relationship between scientists and the public has a specific characteristic: scientists cannot know immediately whether the evidence thresholds they set are the thresholds that will avoid a certain social negative impact or will promote a social positive impact (Leonard and Tochia, 2022; Roberson et al., 2022). They need the public to confirm whether a certain threshold preserves the public wellbeing. For this reason, scientists cannot know by themselves whether their threshold setting matches public interests.

To illustrate why this situation of scientific uncertainty regarding the match between scientific and public interests arises in autonomous systems projects, we describe two projects analyzed by two RRI practitioners who are promoters of the participatory approach. The first project developed Athena AI (Roberson et al., 2022). Athena AI is a decision support system that, through different types of sensors (visual, thermal, among others) displays information for military operators about a target on the battlefield. This autonomous system provides information regarding the impact point on the target, the effects on using a specific weapon on the target, and the incidental damage to nearby people, vehicles, buildings and infrastructure (Roberson et al., 2022, p. 2). Roberson et. al. (2022) indicate that Athena AI was developed for reducing the fog of war, i.e., the uncertainty that military operators have about whether their actions in the battlefield comply with international humanitarian law. Therefore, the experts involved in this project did not know by themselves how to design and implement the autonomous system and, at the same time, comply with this legal and ethical framework.

Following the RRI methodology, these experts involved the stakeholder groups that represent public interests, and, through a series of workshops and ethical and legal guidelines applications, they designed and implemented the autonomous system in accordance with the indications these stakeholders gave about how to protect public interests. This shows that the Athena AI case highlights the importance of involving public agents for setting the evidence thresholds that support the scientific decisions that will preserve public wellbeing. In this way, this case shows that assuring enhanced public trust needs the participation of the public agents in the development of autonomous systems.

The second project had as its goal the development of a qualitative case study on industrial cleaners' attitudes towards autonomous cleaning robots (Leonard and Tochia, 2022). The RRI practitioners who analyzed this case show that these autonomous systems are deployed in situations where there is a power relation between experts and the public. The project evaluated a specific type of autonomous cleaning robot: the UV disinfecting robot Zappy was designed to verify through UV sensors whether a certain surface was effectively disinfected by a human cleaner. The RRI practitioners (who were also involved in researching the cleaners' attitudes to this type of robots) found that the cleaners do not want to give their opinion or their cleaning experience because they have the fear that the management will put them out of work. The function of Zappy is a cobot function. This means that it collaborates with human cleaning staff in cleaning tasks. As cleaners do not have the power to define their working conditions or to assure their work continuity (this is decided by the management of the cleaning agency), they fear of giving the information to the designers of Zappy regarding how this cobot can complement the cleaning tasks. Therefore, these experts do not know how to design Zappy in a way that would ensure the wellbeing of the cleaners. This fear is worsened by the fact that Zappy is a robot, so it is capable of performing physical tasks. Hence, cleaners fear that Zappy will not collaborate with them but will replace them. Thus, this case shows that in autonomous systems work or industrial applications, the power relations between employers and employees impedes that experts know how to design autonomous systems and ensure that their decisions are guided by human values.

Both cases show relationships between experts and public agents where experts cannot know by themselves how to take scientific decisions according to the human values of these agents (Leonard and Tochia, 2022; Roberson et al., 2022). Consequently, both cases show

that to ensure that these decisions are guided by these values it is required to enable channels where these agents can communicate to scientists how to protect these public agents' interests. In the first case, workshops where public stakeholders take place and guidelines produced by these same stakeholders ensure this communication. In the second case, Leonard and Tochia (2022), indicate that it is necessary to create dialogues and spaces between scientists and public agents that make these agents feel safe and assure them that their interests are being protected. For instance, this can be achieved by applying methodologies that make scientists aware of their biases and the power imbalances when elaborating and conducting interviews with the public agents. Thus, according to RRI practitioners proponents of the participatory approach, these methodologies that enable the participation of public stakeholders in scientific or expert decision-making guarantee the preservation of the match between scientific and public interests, and, thereby, forge relationships of public enhanced trust.

The transparent approach

In contrast to the participatory approach, proponents of the transparent approach consider that enabling the participation of public agents in the management of evidence and of inductive risk in autonomous systems projects does not ensure that the values of these agents will determine the thresholds for using evidence to support an acceptance/rejection decision. The proponents of the transparent approach agree with the participatory approach that, if scientists aim to manage inductive risk guided by the values of public agents, they need to involve them in their decision-making tasks. However, the proponents of the transparent approach underline that, in autonomous systems projects, the trust public agents place in scientists is not continuous. Scientists must gain the trust of these

agents and keep this trust during the entire development of the project.

The reason why scientists must gain the trust of public agents can be illustrated by the two following cases where RRI advocates of the transparent approach analyze socially embedded autonomous systems projects. First case: Tekkesinoglu et al. (2025) analyze autonomous vehicles. These RRI practitioners note that the trust relationship between autonomous vehicles and human users can be characterized in the following way: humans trust a machine (the autonomous vehicle) that is not a human but, as it performs human actions (driving) in a human context (streets and roads), it should behave as a human. Thus, every time the autonomous vehicle deviates from human behavior, the human user loses trust in the vehicle and this vehicle must regain it.

Second case: Patel et al. (2024) analyze applications of autonomous systems where humans must follow instructions given by an autonomous system in an unplanned way. For example, an autonomous robot in charge of supporting humans in emergency situations warns the user that he must not follow through a door of the building because there is fire on the other side. Suppose that the robot was never activated before because there were no emergencies until now. Patel et. al. (2024) note that trust requires time for its development, it never happens instantaneously. Also, trust requires that the trustor must have experience of the trustee before she or he is able to trust this trustee. Therefore, when a human user must follow instructions from a robot instantaneously and without having met the robot before, namely, in an unplanned way, this human user passes from a state of non-trust to a state of trust. It is not important to consider in this case that humans do not have the normal resources (time and experience) for developing trust. The relevant point is that,

in these autonomous systems applications, where humans must follow spontaneously the instructions given by a robot, human users must shift from a state of non-trust to a state of trust.

Hence, these two cases show that autonomous systems applications produce relationships between experts and the public where scientists must gain or produce the trust of the public. In the first case, the designers of the autonomous vehicle must develop a mechanism for gaining the confidence of the human user, who is the public agent that must trust that the vehicle is going to keep her/him safe. In the second case, the designers of the autonomous robot must ideate a way for gaining in the short term the trust of a human that has never interacted with this robot before. The significant aspect of these cases is that they demonstrate that producing enhanced public trust does not consist solely of involving public stakeholders. Scientists must also gain the trust of the public.

Previously, we have stated that in the relationships of enhanced public trust three conditions are met: 1) the preservation of public interests depends on the scientists; 2) the public believes that the scientists will protect public interests; 3) the interests of the scientists should match the interests of the public. The first condition is always met in scientific and technological projects embedded in societal contexts, since these projects always have social impacts. Also, we have shown that the fulfilment of the third condition is achieved by enabling public stakeholders to take part in scientific decision-making processes. Additionally, as the fulfilment of the third condition matches the interests of scientists with the interests of the public, the public can assure and believe reasonably that scientists will defend public interests. Hence, the fulfilment of the third condition brings the fulfilment of the second condition. Consequently, if the transparent approach advocates note that in autonomous systems projects public

agents transition from non-trust to trust or from trust to non-trust, there must be some element in these projects that impedes the continuous participation of public stakeholders in scientific decision-making. Thus, if advocates of the transparent approach do not consider that public involvement in scientific activities is enough for achieving public enhanced trust, they consider that it is necessary to find a way to fulfil the second condition.

The analysis of Tekkensinoglu et al. (2025) of autonomous vehicles shows how this condition can be met. These RRI practitioners note that the most important feature autonomous vehicles must have for making users trust them is transparency. Transparent autonomous systems enable users to see that 1) all the data necessary for making a decision based on the users' characteristics and values is being obtained by the system. For example, transparent autonomous vehicles are aware that younger users might not be familiarized with vehicles, so these systems must be more careful when transporting these users, e.g., they must drive at a lower speed. Hence, demographic data, such as age, must be part of the decision-making data. 2) Users can see how this data is transformed into a decision clearly. 3) The relationship between data inputs and decision outputs is logical or reasonable from a human perspective.

Patel et al. (2024) also emphasize the importance of transparency for developing trustworthy autonomous systems. These authors focus on the concept of "sensemaking". Sensemaking is defined as the process where all the variables of a certain situation are transformed into a reasoning process that enables a human to comprehend why these variables lead to a certain decision. In this way, autonomous robots instructing humans in emergency situations are able to communicate to these humans these comprehensible reasoning procedures.

Thus, these analyses show that it is not enough to include public agents in scientific decision-making for assuring enhanced public trust relationships in autonomous systems projects. Instead, it is also necessary to make these systems transparent. Therefore, it is necessary that the users of autonomous systems are capable of seeing clearly how their values lead these systems to certain behaviors and decisions. Thus, the transparent approach adds a second requirement for enhanced public trust. The first requirement is the involvement of public agents in scientific decision-making. The second requirement is the capacity of autonomous systems of showing how human values produced specific decision outcomes.

5. Cost-sensitive Optimization as an Approach for achieving Enhanced Public Trust

We consider that the transparent approach is problematic. The reason for this is that this approach does not consider the *numerical uncertainty* associated with accepting/rejecting mathematical modelling solutions. As we indicate in the second section, autonomous systems are supported by ML models and, thereby, by mathematical models. Therefore, proponents of the transparent approach omit a central aspect of autonomous systems in their analyses. We note that this omission leads them to consider that the way experts achieve transparency in autonomous systems is by applying a *rule-based* approach. This means that they consider that an autonomous system is transparent when it can convey to users the following reasoning procedure:

- (1) User A defends the value X
- (2) Situating the threshold a at the level x will protect X

(3) Therefore, the scientist situates the threshold a at the level x and, consequently, protects X .

For example,

- (1) User A wants to feel safe while riding an autonomous system.
- (2) Establishing that, just when there is 90% probability that the fog will dissipate in the road, the autonomous vehicle will not turn on the dangerous weather alarm will assure that the user A will not use her vehicle in a dangerous weather.
- (3) Therefore, designers will program the autonomous system to turn on the dangerous weather alarm unless there is 90% probability that the fog will dissipate. Consequently, it assures the safety of A.

Notwithstanding, we consider that conveying this rule-based reasoning procedure is not enough to achieve transparency in autonomous systems. The reason for this is that managing inductive risk in ML modelling cannot be achieved by a rule-based reasoning procedure. When there is a hypothesis or a model to be accepted/rejected based on a human value, four different scenarios might occur: 1) accepting/rejecting a hypothesis/model with a high probability of being true and highly favorable according to this human value (we will call these hypotheses highly favorable hypotheses with *high utility*); in other words, scientists produce a true positive; 2) accepting a hypothesis with a low probability of being true, but with high utility (false positive); 3) accepting a hypothesis with high probability but with low utility (true negative); 4) accepting a hypothesis with low probability and with low utility (false

negative).⁴ Therefore, depending on the position where the scientist situates the evidence strongness threshold, one of the four outcomes is achieved.

The important element to consider is that, when situating this threshold, there are two desirable outcomes: accepting a hypothesis with high probability and with high utility, and accepting a hypothesis with low probability and with high utility. Also, there are two undesirable outcomes: accepting a hypothesis with high probability and with low utility, and accepting a hypothesis with low probability and low utility. Additionally, setting the threshold in whichever point increases the likelihood of two outcomes and the inductive risks associated with these outcomes, not one outcome. For example, situating the threshold in a high point will decrease the likelihood of a hypothesis with low probability and with high utility. Also, it will decrease the likelihood of a hypothesis with low probability and with low utility. For this reason, scientists cannot reason in a rule-based way, where each action, according to a rule, leads to a single desirable result; scientists cannot act considering that, if the threshold is situated at a certain level, they will obtain a specific single outcome. Also, when situating the threshold, scientists must prioritize which outcomes are preferable to obtain or to avoid, and balance these utilities with the probabilities of these outcomes. Consequently, it is not enough that a human user indicates to a

⁴ Every acceptance decision entails a rejection decision. For example, when scientist accept “the drug is toxic” they reject simultaneously that the drug is not-toxic. For this reason, there are always four possible results in scientific acceptance/rejection of hypotheses or models: two true results (true acceptance and true rejection, i.e., true negatives and true positives) and two false results (false acceptance and false rejection, i.e., false negatives and false positives. These results can be displayed in a 4x4 cost matrix, where scientists and other stakeholders estimate a cost for each of the four results. For more information see (Karaca, 2021, p. 102).

scientist the value she defends to manage inductive risk and to situate evidence strongness thresholds guided by this value. Scientists, together with users, should contemplate the multiple outcomes and the multiple options for locating thresholds and make the decision where to situate these thresholds based on the *balance* between the *utilities* and *probabilities* of these outcomes. Only the exercise of making this balance will guarantee full transparency in autonomous systems and, thus, the adequate involvement of public stakeholders in scientific decision-making performed in the development of these systems. In fact, public stakeholders will see that managing inductive risk is not a rule-based reasoning procedure, but a process of assigning utilities to outcomes and balancing them with these outcomes' probabilities.

We argue that ML cost-sensitive optimization is a process that enables this balance between probabilities and utilities. As we mentioned in the second section of this article, in this process, scientists assign a cost to each modelling solution based on the human values of the social context where the ML application is developed. As the learning algorithm finds the probabilities of the modelling solution, if each solution has a cost or a utility, the learning algorithm will find the solution with the optimal balance between probability and utility (cost). In this way, scientists together with public agents must assign costs to modelling solutions and optimize the modelling according to this cost assignment. For this reason, we defend that the only way of achieving enhanced public trust in autonomous systems projects is by applying cost-sensitive optimization processes in the ML models that support the operations of these systems.

6. Discussion

In this last section, we evaluate two possible counterarguments to our proposal of using cost-sensitive optimization for achieving public

enhanced trust in autonomous systems. First, we articulate the argument and, second, we give our reply.

(A) The rule-based approach is necessary in some decision-making scenarios of autonomous systems projects

A possible counterargument to our proposal is that there are scenarios of autonomous systems projects where a rule-based approach is necessary. Suppose that an autonomous vehicle user had to take an alternative route to his house in the countryside because the main route is under maintenance. Also, suppose that this alternative route is covered by snow, so it is slippery. Now, at some point of the route, a deer crosses the road unexpectedly. Normally, braking decision systems supported by ML are optimized in a way that false positives are more favorable than false negatives. It is preferable to break even if an object was falsely detected than to continue even if an object was not detected. However, this situation is different, since braking in a slippery road can overturn the vehicle and put the passenger at serious risk. The designers of the autonomous vehicle could not have optimized cost-sensitively the model, since this situation is unexpected. In this scenario, it is preferable to establish a rule that assures safety to the passenger and other vulnerable beings such as the deer. For example, designers could establish that, if the road is slippery and the system detects an object straight ahead, the system makes an evasive maneuver.

Thus, for achieving transparency, the designers indicate in the user guide that, promoting the value of safety, the vehicle will always react in this manner when the conditions of slippery road and object detected are met. Therefore, the transparent approach proponents could say that, in autonomous driving projects, to achieve enhanced public trust in autonomous vehicles, public agents could be invited to decide the vehicle rules of action and reaction based on their values,

such as safety. In this way, no cost-sensitive optimization is needed, and the transparency condition is met. We consider that this unexpected situation can be addressed by a cost-sensitive approach. The autonomous vehicle could be connected to a geographic information system (GIS) that continuously actualizes the weather information and how the weather information might affect autonomous driving. Therefore, this situation could be calculated during the design of the autonomous system and, consequently, it could be incorporated into the cost-sensitive optimization process. Moreover, the braking system could be designed in a way that slippery roads do not cause rollovers when the vehicle brakes unexpectedly.

Thus, we establish that autonomous systems projects stakeholders do not need to use at any moment a rule-based approach. In cases where it is not clear which decision the vehicle should make, stakeholders apply a cost-sensitive optimization process. In cases where it is not possible to apply this process, stakeholders introduce measures that do not require decision-making. Following the previous example, if it is possible to connect the autonomous vehicle to the GIS system, then stakeholders perform cost-sensitive optimization. If it is not possible, then stakeholders redesign the braking system and other features of the system, with the aim of avoiding this unexpected situation from putting any stakeholder at risk.

(B) Cost assignment cannot be done in ML modelling solutions that model environmental data

A second possible counterargument asserts that it is not possible to assign costs to modelling solutions that model environmental data. In a credit scoring system, it is clear, or at least plausible, that people with short credit history have high probability of entering in default. Therefore, it is clear that a decision that grants a credit to a person with

short credit history has a high cost for a credit institution that defends financial security and a low cost for an institution that defends financial risk-taking. Therefore, in this type of ML application that does not model environmental data, but financial data, it is possible to assign costs to modelling solutions. By contrast, it is not possible to assign costs to solutions that model environmental data. For example, how to assign costs to a modelling solution that classifies the weather as foggy?

One stakeholder might answer that a solution that classifies the weather as foggy has a high cost for autonomous vehicles stakeholders that defend users' safety. Namely, this solution will discourage people from using their autonomous vehicles. In this way, they will be safe because they will not be driving in adverse environmental conditions. However, another stakeholder might answer that foggy weather does not discourage people from using their vehicles. Rather, users might activate safe modes of driving, like driving very slow or less aggressively, or solely using roads that are very safe. Additionally, this safe mode of driving will allow users that must use their vehicles to use them despite the weather conditions. If a user must go to the hospital in her vehicle and the vehicle drives in a very safe mode, perhaps it is safer for this user to take a ride to the hospital than staying at home. Thus, this second stakeholder does not consider that a classification of the weather as foggy has a high cost for stakeholders who promote safety; both accepting and rejecting that the weather will be foggy can impact safety at the same degree. If cost assignments cannot be done, then neither can cost-sensitive optimization.

We reply to this counterargument by stating that stakeholders can establish conventions for assigning costs to environmental data modelling solutions. Following the previous example, stakeholders

can analyze previous statistical studies that compare the rate of automobile accidents in foggy weathers vs the rate of automobile accidents in non-foggy weathers. If the rate of accidents in foggy weather increases significantly, then it is reasonable for stakeholders promoting safety to convene that modelling solutions that classify weather conditions as foggy have a lower cost than those solutions that do not classify weather conditions as foggy. Thus, with these convention mechanisms, autonomous systems stakeholders can assign costs to environmental data and perform cost-sensitive optimization.

7. Conclusion

In this chapter, we have examined the relationship between inductive risk in autonomous systems and how AI Ethics can be applied in the development of these systems with the aim of guiding decision-making based on human values. We have pointed out that an aspect of great importance for ethical decision-making is how to ensure an enhanced trust relationship between the experts who manage the inductive risk arising in these systems and the public agents impacted by these systems. Until now, two major approaches have been proposed for forging these trust relationships: the participatory approach and the transparent approach. We argued that none of these approaches adequately addresses this problem of enhanced public trust. Therefore, we proposed a third approach, the cost-sensitive approach, that highlights the importance of involving public stakeholders in the process of assigning costs and optimizing the modelling solutions produced by the ML models that support the functioning of autonomous systems. After the exposition of our approach, we addressed two possible objections.

This chapter has three main contributions. In the first place, it shows that cost-sensitive optimization is a process in ML applications that is

not limited to application cases where social values must be taken into consideration, but also to cases where the public stakeholders must take part in ML modelling optimization, such as autonomous systems. Therefore, it reveals that the possibilities of applying this process are wider than those conceived until now. In the second place, we show that autonomous systems, due to their functioning through ML modelling, are a type of technological application that requires a deeper involvement of public stakeholders than other scientific or technological applications. The RRI approaches used until now for addressing the participation of public agents are limited because they do not consider the need to involve these agents in ML models cost-sensitive optimization. In the third place, by addressing the objections, we show that the fact that autonomous systems work principally with environmental data and are pervaded with unexpected situations is not an obstacle for performing cost-sensitive optimization in these systems.

A central topic in the RRI methodology is that its application consists not only of formulating the actions that attain enhanced public trust, but also in answering how to perform these actions. For example, Roberson et al. (2022), advocates of the participatory approach, indicate that, to involve public stakeholders in the development of Athena AI, they organized a series of workshops and utilize ethical and legal guidelines that enabled these stakeholders to indicate how to design and implement this autonomous system according to their human values. Also, Tekkensinoglu et al. (2025, p. 7), establish that it is necessary to use user-centric explanations, i.e. comprehensible and meaningful explanations, for conveying transparently to public stakeholders the reasoning behind the decision-making processes performed by autonomous systems. Our assessment of the participatory and transparent approaches has focused on the actions of these approaches: involving public stakeholders and making

transparent the reasoning of autonomous systems' decision making, respectively. Thus, we have proposed a new action for achieving enhanced public trust: performing cost-sensitive optimization with the participation of the public stakeholders. Nonetheless, we have not considered how to accomplish this public cost-sensitive optimization. For this reason, we consider that a topic for further research could be how to attain or operationalize this type of cost-sensitive optimization.

References

- Biddle, J. (2022). "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning", *Canadian Journal of Philosophy*, 52(3): 321-341.
- Carrier, M., & Irzik, G. (2021). Responsible research and innovation: Coming to grips with an ambitious concept. *Synthese*, 198(Suppl 19): 4627-4633.
- Douglas, H. (2000). "Inductive Risk and Values in Science", *Philosophy of Science*, 67(4): 559-579.
- Elkan, C.P. (2001). "The Foundations of Cost-Sensitive Learning" In *The Proceedings of International Joint Conference on Artificial Intelligence*, pp. 973-978.
- Fernández, J., Perez, J., Agirre, I., Allende, I., Abella, J., & Cazorla, F. J. (2021). "Towards functional safety compliance of matrix-matrix multiplication for machine learning-based autonomous systems"; *Journal of Systems Architecture*, 121. <https://doi.org/10.1016/j.sysarc.2021.102298>

- Irzik, G., & Kurtulmus, F. (2021). "Well-ordered science and public trust in science", *Synthese*, 198 (Suppl 19): 4731–4748.
<https://doi.org/10.1007/s11229-018-02022-7>
- Jedličková, A. (2024). "Ethical approaches in designing autonomous and intelligent systems: A comprehensive survey towards responsible development", *AI & Society*. <https://doi.org/10.1007/s00146-024-02040-9>
- Karaca, K. (2021). "Values and inductive risk in machine learning modelling: the case of binary classification models", *European journal for philosophy of science*, 11(4); 102. <https://doi.org/10.1007/s13194-021-00405-1>
- Leikas, J., Koivisto, R., & Gotcheva, N. (2019). "Ethical Framework for Designing Autonomous Intelligent Systems", *Journal of Open Innovation: Technology, Market, and Complexity*, 5(1): 18.
<https://doi.org/10.3390/joitmc5010018>
- Leonard, P., & Tochia, C. (2022). "From episteme to techne: Crafting responsible innovation in trustworthy autonomous systems research practice"; *Journal of Responsible Technology*, 11(10): 100035.
 (doi:10.1016/j.jrt.2022.100035).
- Patel, M., Weerawardhana, S., & Moreau, L. (2024). "Jointly navigating through RRI in practice: lessons from technology design.", *Journal of Responsible Innovation*, 11(1).
<https://doi.org/10.1080/23299460.2024.2425091>
- Roberson, Tara & Bornstein, Stephen & Liivoja, Rain & Ng, Simon & Scholz, Jason & Devitt, Kate. (2022). "A method for ethical AI in Defence: A case study on developing trustworthy autonomous systems". 10.48550/arXiv.2206.10769.
- Roberson, T., Bornstein, S., Liivoja, R., Ng, S., Scholz, J., & Devitt, K. (2022). "A method for ethical AI in Defence: A case study on developing trustworthy autonomous systems".
 10.48550/arXiv.2206.10769.

- Sharma, R. (2021). "Big Data for Autonomous Vehicles" In Ahmed, K.R., Hassanien, A.E. (eds.) *Deep Learning and Big Data for Intelligent Transportation. Studies in Computational Intelligence, vol 945*. Cham: Springer.
- Su, K., Zhao, Q., Tomioka, Y., & Liu, Y. (2024). "Cost-Sensitive Road Obstacle Detection for Low-Cost Autonomous Vehicles" In *2024 IEEE Intelligent Mobile Computing (MobileCloud)* (pp. 24-31). Shanghai, China.
- Tekkesinoglu, S., Habibovic, A., & Kunze, L. (2024). "Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap". *ArXiv, abs/2404.00019*.
- Winsberg, E. (2012). "Values and uncertainties in the predictions of global climate models", *Kennedy Institute of Ethics journal*, 22(2): 111–137.

Chapter 10

Neurotechnologies and New Processes of Subjectivation: Problematizing Brain-Computer Interfaces from Michel Foucault's Perspective

Luiza Cunha¹ and Marcela Castanheira²

In this chapter, we aim to point out some philosophical challenges in the context of experiments involving the neurotechnology approaches known as Brain-Computer Interface (BCI). Initially, we present the functioning of BCIs, illustrated with by a hypothetical case study. Then, we use Michel Foucault's texts on power relations in the processes of subjectivity formation as a theoretical reference to address the emergence of the “cerebral subject”. Finally, we discuss and try to contribute to the analysis of some ethical and philosophical dilemmas raised by the use of these neurotechnologies.

Keywords: neurotechnologies, Brain-Computer Interface, modes of subjectivation, brain.

¹ Graduated in Law and has a PhD in Law and Political Science at the University of Barcelona, line of research in Bioethics and Law.

² Graduated and has a master's degree in Philosophy at the Federal University of Goiás. Has a PhD in Philosophy at the Federal University of Minas Gerais.

1. Introduction

The advancement of technologies interfacing with the human brain is progressing exponentially and efficiently. From a historical perspective, research to understand the human brain has advanced significantly. A notable milestone occurred in 2013 with the Brain Initiative (The Brain Research Through Advancing Innovative Neurotechnologies), launched by then U.S. President Barack Obama, investing a high budget to accelerate the development and application of new brain technologies (Bargmann C.I.; Newsome, W.T., 2014).

Neurotechnologies have advanced in seemingly contradictory directions, since they can be used for curing and predicting brain-related diseases, but also may be used for manipulating ideas, behaviors, decisions, and emotions.

While neurotechnology has made remarkable advances in diagnosing and treating neurological diseases, enhancing human life expectancy, its potential for manipulating human behavior, decisions, and emotions opens a complex field of ethical problematization³, once it enables the formation of subjects that are increasingly obedient to the productivist logic of capital.

³ In the interview he gave to P. Rabinow on May of 1984, titled *Polemics, Politics and Problematizations*, Foucault says: "I thought there was an element that could characterize the history of thought: it was what you would call problems or, to be more exact, problematizations. What distinguishes the thought is that it's totally different from the attitude field that might determine it. The thought is not what's present in a demeanor and gives it meaning; it is, most importantly, what allows one to take a distance from the way to do and react and take it as an object of thought and question it about its meaning, its conditions, its purposes. The thought is freedom in relation to what is done, the movement from which we separate, constitute it as an object and think it as a problem" (Foucault, 1994, p.597).

The discussion around this duality may give us the feeling that the advancement of neurotechnology, due to its benefits for the maintenance and conservation of human life, is not harmful. This is the case, for example, of neurotechnologies that can help people with neurological disorders to recover or improve their cognitive capabilities, even helping them with both overcoming addictions and changing harmful behaviors. Such interventions, when carried out ethically and respectfully, with informed consent and in the better interests of the person, can be considered positive. This advancement raises questions about the authenticity of our choices and the integrity of our selfhood, especially when neurotechnologies are combined with artificial intelligence. The promises of healing and enhancement brought by BCIs and neuromodulation devices alter our self-conceptions and potential selves.

We are facing a complex relation between progress and ethics in neurotechnologies. New ethical, axiological, and consequently legal issues arise. The search for ethical balance demands clear criteria for continuous negotiation between technological advancement and ethical principles such as transparency, informed consent, regulation, privacy, accessibility, and promotion of autonomy. The discussion should consider if neurotechnologies can ethically and responsibly serve humanity, questioning not only if they can improve quality of life without compromising our essential freedom and individuality, hence what kind of life⁴ they promote and for whom.

As we dive deeper into the era of neurotechnologies, we confront an old paradox: technological advancement promising to revolutionize

⁴ Foucault states that “a way of life can be shared by individuals of age, status and different social activities. It can make room for intense relationships that don’t seem like none of those that are institutionalized and it seems to me that a way of life can make room for culture and ethics” (Foucault, 2004, p. 165).

the treatment of neurological diseases also brings the possibility of fundamentally transforming human experiences into controllable, predictable, and insensible objects.⁵ Neurotechnology, including BCIs and neuromodulation devices is fundamentally premised on the idea that interfering with our brain can determine who we are by influencing our thoughts, behaviors, and decisions.

As this examination of the intersection between neurotechnology, artificial intelligence, and human experiences concludes, we glimpse a future filled with both emancipatory and oppressive potential. While the advancement of these technologies may be part of the effort to understand and improve the human condition, it raises pressing ethical questions: to what extent is brain intervention, the concrete and palpable foundation of ourselves, ethically acceptable? Do the behavioral changes induced by these technologies alter our self-understanding, i.e., how we think about who we are and who we want to be? Will the advancement of these neurotechnologies promote more inclusion, or instead, increasingly exclude those who do not exhibit socially acceptable and desirable behaviors? Finally, what do we expect from these new technologies, what kind of society, values, human integration, and community?

In this chapter, we will discuss some of these issues by first presenting the concepts involving Brain-Computer Interfaces for Parkinson's disease treatment through a hypothetical case, indicating its interaction with artificial intelligence. Then, using Michel Foucault's texts on power relations in the constitution of subjectivities as a theoretical reference, we address the emergence of a new type of

⁵ In that regard, it's worth checking out what Hannah Arendt has to say about the unpredictability of action. To her, it's the contingency (with all its dangers) that allows the new to show up. Cf. (Adverse, H., 2018).

subject, the “cerebral subject”. Finally, we address some ethical dilemmas posed by the use of these new technologies.

2. Case Study - Usage Modes of Brain-Computer Interfaces

To develop our research, we start with the analysis of a hypothetical case from Jonathan Pugh, Laurie Pycroft, and others, in their article “Brainjacking in deep brain stimulation and autonomy” (Pugh et al., 2018). They discuss ethical issues related to mental privacy, physical and psychological integrity of the individual, in light of the possibility of a third party hacking an individual's neural circuits, raising questions about individual autonomy.

Here is the hypothetical case we chose for our research:

Carl suffers from Parkinson’s Disease, and has validly consented to undergo DBS in order to ameliorate his motor impairment. Although the stimulation is successful in reducing his tremor, Carl is one of the rare patients who experiences off-target effects following stimulation. Whilst undergoing stimulation, he exhibits hypersexual behavior that he does not exhibit in the absence of stimulation. Carl understands this, and chooses to stimulate himself only in the safety of his own home. However, a malevolent third party intentionally initiates stimulation whilst Carl is running errands in town, and Carl commits an act of sexual harassment. It quickly becomes clear to Carl’s treatment team and the police that Carl’s implant had been hacked at the time of his action. However, Carl does not feel substantially different now his stimulation has ceased. *He begins to wonder whether he is his ‘right self’ now, and whether he can ever be regarded as responsible for his actions when there is always the possibility that his brain implant has been hacked.* (Pugh et al., 2018)

From this case, where Carl feels extremely embarrassed for having committed a sexual act despite of being aware of the risk and thus only stimulating himself in a controlled environment, we see both the functioning of a BCI and the harm of brainjacking during the tool's use in treating a neurological disease.

This case raises several ethical problems, like the misuse of the device by others and the effect of hypersexualization. For us, the most important issue highlighted is Carl's self-questioning post-assault: his identity and responsibility for his actions.⁶ He ponders whether his existence and identity can be reduced to cerebral stimuli from the device. Can he be responsible for future actions not knowing if he was influenced by the device? Is he autonomous? How does he understand himself now?

Henceforth, we will discuss important concepts for a better understanding of our case study. It's pertinent to clarify what Brain-Computer Interface and Deep Brain Stimulation are; their correlation with artificial intelligence, and their use in Parkinson's disease treatment.

2.1 Neurotechnologies, Brain-Computer Interface and the treatment of Parkinson's Disease

Neurotechnology, a broad term encompassing technologies for understanding, repairing, replacing, or enhancing neural functions, includes diverse tools like brain pacemakers, Brain-Computer Interfaces (BCIs), and neuroimaging techniques such as Functional

⁶ This matter of penal responsibility makes room for a series of researches about the theme. The person who commits a crime while having their neural system be manipulated through a device implanted in their brain is responsible by their felony entirely? To read more about this subject: Cf. (Thompson, 2021).

Magnetic Resonance Imaging (fMRI). These technologies are transforming our understanding of the brain and opening new possibilities for treating neurological diseases.

One of the current neurotechnologies is Brain-Computer Interfaces (BCIs), which can be conceptualized as a system which allows direct communication between the human or animal brain and an external device. BCIs, integrate neurophysiology, computer science, and engineering. They translate neural impulses into computer commands, striving to link human brains to artificial equations and control external devices (Lebedev and Nicolelis, 2017).

An interesting experiment using Brain-Computer Interfaces involves the implantation of a brain chip in a human being. The chip is capable of processing direct signals that are decoded by an app. Specific chip implantation technology is a rapidly growing area of research, with several companies already deploying similar devices.⁷

The recent progression of Neuralink into human trials, marked by Elon Musk's announcement of a successful brain chip implant in a paralyzed individual, catalyzes a profound philosophical discourse. This event not only signifies a technological breakthrough but also beckons a reevaluation of human-machine integration ethics. The endeavor, while promising therapeutic innovations for debilitating neurological conditions, mandates a vigilant examination of moral, existential, and societal ramifications. Such a leap into intertwining human cognition with artificial intelligence compels a nuanced contemplation of the future of human identity, autonomy, and the ethical boundaries of enhancing human capabilities.

⁷ <https://www.reuters.com/technology/neuralink-implants-brain-chip-first-human-musk-says-2024-01-29/>. Accessed: 30/01/2024.

Furthermore, the transition from preliminary trials on monkeys to the first human patient receiving Neuralink's brain chip implant last Sunday marks a significant advancement in neurotechnology. The patient is reported to be recovering well, with initial results indicating promising detection of neuronal spikes. These spikes, as described by the National Institutes of Health, are neuron activities that utilize electrical and chemical signals to transmit information throughout the brain and body.

The United States Food and Drug Administration (FDA) had previously authorized Neuralink to conduct its inaugural human test last year, a critical milestone in the startup's mission to assist patients in overcoming paralysis and a range of neurological conditions. In September, Neuralink announced it had received approval to recruit for the human trial. This study employs a robot to surgically place a brain-computer interface (BCI) implant in a brain region that controls the intention to move, aiming initially to enable individuals to control a computer cursor or keyboard solely with their thoughts. The implant's "ultrafine" wires are designed to facilitate signal transmission within the participants' brains.

Neuralink has faced scrutiny regarding its safety protocols, especially after veterinary records revealed that implants in monkeys had led to paralysis, seizures, and brain swelling. Nevertheless, Musk has stated, "no monkey has died as a result of a Neuralink implant," adding that the company selected "terminal" monkeys for the experiments to minimize risks to healthy ones. This deliberate approach underscores Neuralink's commitment to advancing neurotechnology while addressing ethical considerations surrounding animal and human testing.

Another interesting example of BCI use was during the 2014 World Cup opening in Brazil, where a paraplegic person kicked a ball

towards the goal using mind-controlled devices that translated thought into movement commands, activating a mechanical leg.⁸

Extensive studies on BCIs and paraplegic mobility by neuroscientist Miguel Nicolelis and his team have developed a kind of "exoskeleton", allowing paraplegics to move. This mechanical suit, controlled by patients' mental commands through BCI technology, exemplifies neurotechnology's potential (Nicolelis, 2022).

For our research, understanding how BCIs are used in neurological disease treatment and how they might influence actions, sometimes undetectably, is crucial. One potential phenomenon is "brainjacking", the unauthorized control of someone's electronic brain implant (Pugh et al., 2018). The feasibility of hacking BCIs has been proven in both experimental settings and real life. In our case study, a brainjacking action occurred in a circuit called Deep Brain Stimulation.

Deep Brain Stimulation (DBS) involves surgically implanting electrodes into specific brain areas, allowing physicians to alter neural function and thus modify behaviors, actions, and cognitive functions. DBS successfully treats conditions like Parkinson's disease, dystonia, and epilepsy, and experimental treatments are underway for other conditions like clinical depression and anorexia, in addition to corrections of abnormalities and moral behaviors (Pugh et al., 2018).

With advancements in artificial intelligence, Parkinson's disease treatment is being implemented using Artificial intelligence methods (AI). AI is "the application of advanced analysis and logic-based techniques, including machine learning, to interpret events and to support, augment and automate decisions and actions" (Limousin and

⁸<https://g1.globo.com/ciencia-e-saude/noticia/2014/06/jovem-paraplegico-usa-exoesqueleto-chuta-bola-na-abertura-da-copa.html>. Accessed: 29/01/2024.

Akram, 2023). Limousin and Akram's analysis (2023) suggests that while these applications are still in the early stages, they are already widespread and require long-term prospective validation.

AI in DBS over the past twenty years has developed closed-loop or adaptive systems that adjust stimulation according to the patient or their activity. These approaches involve simple algorithms directing stimulation on or off, or at varying intensities, based on recorded signal thresholds (Limousin and Akram, 2023). The results have been excellent for Parkinson's patients, with potential for more complex parameters in the future (Limousin and Akram, 2023). Limousin and Akram, states that *"in the future, machine learning could integrate a broader range of information and deliver more complex parameters of stimulation"* (Limousin and Akram, 2023).⁹

Machine learning, a subfield of AI identifying patterns within data sets (Watts et al., 2020), has been extensively studied by Jeremy Watts, Anahita Khojandi, and others at the University of Tennessee, Knoxville, TN, US. Their research on machine learning in DBS for Parkinson's disease shows promising results (Watts et al., 2020).

Machine learning in DBS for Parkinson's treatment includes: "(1) DBS candidate selection; (2) programming optimization; (3) surgical targeting; and (4) insights into DBS mechanisms" (Watts et al., 2020). For our research purposes, it's sufficient to understand the various

⁹ A super interesting use of the machine learning at DBS can be called "image enhancement or 'quality transfer'". In that case, the devices are being utilized to train deep learning algorithms this way: "high-quality MRI datasets, for example, from ultra-high-field MRI, have been used to train deep learning algorithms to identify noise, distortion and artefacts in lower-quality MRI datasets, resulting in images with increased resolution and signal-to-noise and contrast-to-noise ratios. This approach allows direct visualization of DBS targets using conventional, clinically available MRI systems with all the image-quality advantages of higher-field research scanners" (Limousin and Akram, 2023).

ways AI can be connected to Brain-Computer Interfaces. In Parkinson's treatment, AI is used for symptom detection, disease progression prediction, patient response to medication using motor symptoms, optimizing clinical evaluations, and medication regime planning (Watts et al., 2020).

Ultimately, these are just examples of how AI can be used for treating neurological diseases. Our research is particularly interested in demonstrating the potential of neurotechnologies to influence our consciousness, with the real possibility of a third party deciding which thoughts to engage with at any moment (Zohny et al., 2023), as illustrated by the case of Carl presented above.

3. Processes of Subjectivation in Michel Foucault's Thought

Following our discussion on BCIs, we now will focus on the challenges these discourses and techniques pose, using Michel Foucault's work on the constitution of subjectivities. Our theoretical basis here makes two essential movements: first, presenting Foucault's concept of subjectivation to address the emergence of the "cerebral subject"; then, discussing potential ethical and philosophical dilemmas involving Brain-Computer Interface (BCI) experiences.

Foucault's concept of subjectivation unfolds into various questions in his works, encompassing discourses, powers, and moral actions. Broadly, his early works (1961-1969) question the subject as the founding instance of knowledge; his genealogy of powers (1970-1979) addresses mechanisms of subjugation controlling both individual bodies and populations; and his later texts (1980-1984) focus on ethical

processes of subjectivation, i.e., the relationship with oneself.¹⁰ Although the texts and concepts differ, they illuminate the formation of subjectivities in varied human practices linked to knowledge (discursive and non-discursive practices), power (coercive practices), and self-relations (practices of the self).¹¹

Foucault emphasizes that subjectivation modes are constituted by historically situated practices, meaning the subject emerges as an effect of different historical practices.¹² Subjectivity isn't tied to an essence, substance, identity, or nature, but is a historical figure that's fragile, open, multiple, and unfinished, linked to how one *becomes* a subject through specific human practices.

How do we make sense of human practices forming our subjectivity field? Foucault finds a key in the knowledge-power dyad, noting knowledge exists in domination and submission relationships. Power and knowledge mutually imply each other, as discourses support

¹⁰ Here we are considering the process of subjectivation in its broader sense, as practices that constitute the self. In that regard, the subject may appear as an object of knowledge and power, in a mutual relationship with subjectivation and objectification. In a stricter sense, subjectivation refers to the ethical process of a subject's self-constitution. (Cf. Castro, 2009, p.408).

¹¹ During the Foucauldian path, the practices take three shapes: discursive practices and non-discursive practices, social practices, and practices of the self. Each has a function: discursive practices and non-discursive practices outline the theory of subject and the subject of knowledge; the social practices (punitive, government-related, and juridical) deviate their analysis from a theory of power; at last, the practices of the self stray him further from theories about the subject. (DARDOT, p. 2015). The notion of practices allows Foucault to make a history of subjectivity that does not search for an identity of human nature or the truth hidden in the conscience's secrets. In this article, we would like to think of how new practices involving Brain-Computer Interface could shape a new type of subject, a cerebral subject.

¹² It's worth mentioning here that Foucault's distancing from Cogito's philosophical tradition, which admits the subject as a thinking substance that is finished, not open to transformations, but is ready to learn.

power techniques, which are knowledge production mechanisms.¹³ Considering Foucault's knowledge-power circularity, we hypothesize¹⁴ that neuroscience and neurotechnology, developed and refined in the 20th and 21st centuries, creates a new discourse and power mechanism set, producing a specific subjectivation mode: the "cerebral subject" (Azambuja, 2017, p. 26).

The "cerebral subject" concept, arising from neuroscience and neurotechnology, inaugurates practices, discourses, and forms of thought where the brain is central to subjective constitution, whether healthy or diseased. It equates the subject with the brain: understanding the brain equals understanding the person. This redefines being human based on cerebral characteristics¹⁵, positioning the brain as the unequivocal truth bearer about the subject. This truth, translatable into digital information, can be manipulated for

¹³ In that regard, it's worth checking out Foucault's studies about prison and psychiatric power. Although they are different institutions, both illustrate clearly the circularity between knowledge-power: as the individual is controlled and watched inside disciplinary institutions, it is produced knowledge about him. Cf. (Foucault, 2006); (Foucault, 2013).

¹⁴ The experiments executed by the renowned Brazilian neuroscientist Miguel Nicolelis were a source of inspiration for this hypothesis. He said, after the result of his research, the following sentence: "The brain was, finally, free from the body and could act over the world directly, by only producing what it produces every second: electrical activity that could now be utilized to create movement. The brain didn't need the body anymore." Likewise, that sentence needs to be analyzed, because it is sensible to question how someone could say that "the brain didn't need the body anymore" and make this declaration sound so self-evident? The brain doesn't need the body anymore. It creates its own. (Azambuja, 2017, p. 128). Here the problem with the relation between body and subjectivity is laid out. And in that regard, the brain is not the biological organ, but a device. Cf. (Azambuja, 2017).

¹⁵ It's worth mentioning that as a counterpoint to this concept, other concepts of being human can be found in philosophical, religious and mystical tradition.

corrections and adjustments, as in the previously mentioned case of Carl.

Rather than questioning the brain's importance, we must consider how its functioning suffices for human action involving moral choices, mental suffering, sexual practices, etc. The “cerebralization” of behaviors, where the brain becomes the exclusive mental state agent, offers mechanical, reductionist explanations for human phenomena like mental disorders, behaviors, and actions (Zorzanelli, R.; Ortega, F., 2011, p.32).

Azambuja (2017, p. 41) suggests equating the brain's place in neuroscience discourses with sexuality's role in *scientia sexualis*.¹⁶ Neuroscience mechanisms intervening in human behavior and action

¹⁶ The expression *scientia sexualis* is used by Foucault in the first volume of *History of Sexuality* to outline an opposition with *ars erotica*. While this one, more common in China, Japan, India, Rome, Arabic Muslim nations, all of them found the truth about sex of pleasure itself “faced as practice and received with experience”, that is, being considered in itself its intensity, quality, duration, reverberations in the body and soul, the first developed ways of knowledge-power on the way of confession to tell the truth about sex. (Foucault, 1988, p. 57). As he states, “More precisely, it was assigned the task to tell truthful discourse about sex, and that is trying to adjust, not without difficulty, the old procedure of confession to the rules of scientific discourse. The *scientia sexualis*, developed in the XIX century, paradoxically, has at its core the singular rite of obligatory and tiring confession, that constituted, in the Christian West, the first technique to find out the truth about sex” (Foucault, 1988, p. 66). It would be necessary to trace a fundamental distinction that is not taken into account in Azambuja’s article: the truth of sexuality was extracted through a hermeneutic process of interior analysis that is done by the subject themselves; when we refer to neurotechnologies, telling truth about someone is up to somebody else, the one who’s capable of reading the functioning of your brain, that is, the neuroscientist - and that ends up being a new way of objectifying the subject through scientific discourse. It’s another way of making yourself visible to the other. We could still establish a correlation between drugs and Brain Computer Interface interventions.

correlate with Foucault's discussions on scientific discourse emergence on sexuality. In the last case mentioned, discourses and power mechanisms focus on sexuality to control individuals and populations. In neuroscience, techniques and discourses focus on the brain as an identity equivalent, decoding it to uncover secrets.

Considering the brain as a subjectivation policy "marker creates a new existence, a new meaning and sense network, new ways to behave, think, and govern humans and the world. It's a new societal organization model, shaping economy, laws, life and death discourses, health care, and disease perspectives" (Azambuja, 2017, p. 40).

Linking the subject with the brain redefines normality, health, and pathology, as the brain becomes life's principle. Neuroscience and neurotechnologies establish a new regime of truth constituting a new subject. We must question this subject form, as neuroscience isn't just a science; it serves economic, political, educational, and social interests beyond individual dimensions, participating in a social model where brain manipulation extends to social, environmental, biological, and individual experiences (Azambuja, 2017, p. 156). The advancement of these neurotechnologies may serve economic and political interests, classifying and segmenting individuals based on pre-formulated judgments about social acceptability and desirability.

One question that arises when we refer to the use of Brain-Computer Interfaces (BCI) is how they affect the way we experience ourselves. Although the use of neurotechnologies like BCIs intensifies what other mechanisms, such as disciplines and drugs¹⁷, already did, the

¹⁷ Discipline is a power mechanism developed mainly during the XVIII century, that affects individual bodies in order to control gestures and behaviors. It is the way power functions inside modern institutions like prison, school, factories and

aggressiveness of their intervention amplifies the unbridled attempt to control human actions. As we have seen, these brain technologies alter our relationship with our body not only in their ability to diagnose, prevent, treat, and cure diseases, thereby prolonging life and optimizing behavior, but primarily because they reduce it to cerebral capacities. The body, far from being a source of errors, desires, actions, impulses, becomes an object manipulable by cerebral interventions.

4. Conclusions

As we have seen, the emergence of Brain-Computer Interfaces (BCIs) and neuromodulation devices, like deep brain stimulators (DBS), have the potential not only to heal but also to reshape our cognition and behavior. They constitute a power-knowledge mechanism that has led to the emergence of a new type of subject. Our goal was to expose some philosophical problems raised by the equalization of subject and brain, especially the philosophical and ethical implications concerning how we conceive ourselves. Our aim was to show that the association between subject and brain could only be made through the discourses, techniques, and practices of neuroscience, suggesting that other ways of thinking about what we are possible.¹⁸

In the hypothetical case studied, for example, we see straightforward contours of a violation of Carl's individual autonomy, to the extent he no longer knows for sure who actually is the person who committed the sexual crime. The most dangerous issue here is the fact that he puts his authenticity, his notion of autonomy, and the way he relates to himself into a conflict. It is a way of acting and thinking in which one

psychiatric hospitals. Both the discipline and psychiatric medication are mechanisms that are used to control individuals and fit them into the norm.

¹⁸ In that regard, it's worth checking out the work of Ortega and Vidal about the history of the cerebral subject in contemporary culture (Ortega, F., Vidal, F. 2007).

does not know what to expect from one's own conduct. This brings relevant debates to the philosophical and ethical field, as the being cannot recognize itself in its actions.

Furthermore, human experiences such as disease, death, sexuality, and crime are viewed as objects by an external gaze capable of preventing or controlling them.¹⁹ These experiences are reduced to objects of knowledge, or rather, to a field of analysis conducted externally by those capable of preventing or correcting them. This undermines the emergence of singularity and seems an attempt to avoid the fragilities inherent in human life.²⁰ This forgets that it also nullifies its contingent and unpredictable nature. Even though this fragility is uncomfortable and brings conflicting and unavoidable demands, it also propels us to aspire to our self-transformation as subjects of action. In this sense, not only our freedom of thought is at stake, still as well our own autonomy and protagonism in our lives.

We conclude by emphasizing the need for a robust philosophical debate that accounts for the complexity and contradictions of these new technologies. Ultimately, we explored the current use of neurotechnologies in treating diseases, questioning whether they can advance in a way that preserves and enriches human dignity and freedom.

¹⁹ Here's worth mentioning that Foucault's critiques about the objectification of human experiences like death, madness, crime, and sickness, from the resurging of *psi* sciences that are born from a new way of putting discourse in our daily lives. This new way of discourse consolidates in the formation of new knowledge that operates in categories of science and administration, in search of neutrality and objectivity of the subjects. It's about psychiatry, criminology, and social medicine. (Cf. Foucault, 2006a).

²⁰ About an implicit norm of life, that looks for immortality in the vital process, imposed ever since modernity. (Cf. Foucault, 1988).

Some questions remain open in this article. However, it's worth noting that when philosophy attempts to prescribe solutions or paths, it risks becoming the foundation of truth in the exercise of power; that is, reinforcing the practical and discursive field in which the subject finds the conditions of their subjugation. Conversely, we consider it our duty to continue diagnosing the problems of our times. As Foucault once said, "I believe the ethical-political choice we have to make every day is to determine what is the main danger" (Foucault, 2004, p. 386). This has been our attempt.

References

- Adverse, H. (2018). "Ação e contingência em A Condição Humana de Hannah Arendt", *Princípios, Revista de Filosofia. Natal*, 25(48): 129-156.
- Azambuja, M. (2017). *Da alma para o corpo e corpo para o cérebro: Os rumos da psicologia com as neurociências*. Porto Alegre: ABRAPSO.
- Bargamnn, C., Newsome, W. (2014). "The Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) Initiative and Neurology", *JAMA Neurology*, 71(6): 675–676.
- Castro, E. (2009). *Vocabulário de Foucault*. Belo Horizonte: Autêntica.
- Dardot, P. (2015). *De la praxis aux pratiques*. In Laval, C., Palterinieri, L. and Taylan, F. (eds.) *Marx and Foucault: lectures, usages, confrontations*. Paris: Éditions La Découverte.
- Foucault, M. (2006). *A Hermenêutica do Sujeito*. São Paulo: Martins Fontes.
- Foucault, M. (2004). "A ética do cuidado de si como prática da liberdade." In: *Ética, sexualidade e política*, 264- 287. Rio de Janeiro: Forense Universitária.
- Foucault, M. (1994). *Dits et Écrits IV, 1980-1984*. Paris: Gallimard.

- Foucault, M. (1988). *História da sexualidade I, A vontade de saber*. Tradução de Maria Thereza da Costa Albuquerque e J.A. Guilhon Albuquerque. Rio de Janeiro: Edições Graal.
- Foucault, M. (2006a). *O poder psiquiátrico. Curso dado no Collège de France (1973-1974)* In Lagrange, J. (ed.) São Paulo: Martins Fontes.
- Foucault, M. (2013). *Vigiar e Punir, nascimento da prisão*. Lisboa: Edições 70.
- Foucault, M. (2004). “Verdade e Poder”. In *Microfísica do poder*, pp.1-14. São Paulo: Graal.
- Lebedev, M. and Nicolelis, M. (2017). “Brain-Machine Interfaces: From Basic Science to Neuroprostheses and Neurorehabilitation”, *Physiological Reviews*, 97(2): 767–837.
- Limousin, P. and Akram, H. (2023). “AI and deep brain stimulation: What have we learned?”, *Nature Reviews Neurology*, 19(8): 451–452.
- Nicolelis, M. (2022). “Brain–machine–brain interfaces as the foundation for the next generation of neuroprostheses”, *National Science Review*, 9(10): nwab206. <https://doi.org/10.1093/nsr/nwab206>.
- Ortega, F. and Vidal, F. (2007). “Mapeamento do sujeito cerebral na cultura contemporânea”, *REECIS - R. Eletr. de Com. Inf. Inov. Saúde. Rio de Janeiro*, 1(2): 257-261.
- Pugh, J., Pycroft, L., Sandberg, A., Aziz, T. and Savulescu, J. (2018) “Brainjacking in deep brain stimulation and autonomy”, *Ethics and Information Technology*, 20(3): 219–232.
- Schusler, A. (2020). “Artificial Intelligence and Mind-reading Machines—Towards a Future Techno-Panoptic Singularity” *Postmodern Openings*, 11(4): 334–346.
- Thompson, K. (2021) “Committing crimes with BCIs: How brain-computer interface users can satisfy actus reus and be criminally responsible”; *Neuroethics*, 14(Suppl 3).

- Watts, J., Khojandi, A., Shylo, O. and Ramdhani, R. (2020) "Machine Learning's Application in Deep Brain Stimulation for Parkinson's Disease: A Review", *Brain Sciences*, 10(11); 809.
- Zohny, H., Lyreskog, D., Singh, I. and Savulescu, J. (2023) "The Mystery of Mental Integrity: Clarifying Its Relevance to Neurotechnologies", *Neuroethics*, 16(3): 20. <https://doi.org/10.1007/s12152-023-09525-2>.
- Zorzanelli, R. and Ortega, F. (2011) "Cultura somática, neurociências e subjetividade contemporânea", *Psicologia e Sociedade*, 23 (n. spe.): 30-36. Disponível em: <https://www.scielo.br/j/psoc/a/xZ36Gqh3f4mZMXpJHYVftdd/?format=pdf&lang=pt>.
- Jovem paraplégico usa exoesqueleto e chuta bola na abertura da Copa. G1 (2014) Available: <https://g1.globo.com/ciencia-e-saude/noticia/2014/06/jovem-paraplegico-usa-exoesqueleto-chuta-bola-na-abertura-da-copa.html>. Accessed: 29/01/2024.
- Elon Musk's Neuralink implants brain chip in first human. Reuters,(2024) Available: <https://www.reuters.com/technology/neuralink-implants-brain-chip-first-human-musk-says-2024-01-29/>. Accessed: 30/01/2024.

Chapter 11

Brain-Machine Interfaces and Artificial Intelligence: An Ethical Overview

Steven S. Gouveia¹

This chapter critically examines the ethical implications of integrating Artificial Intelligence (AI) with Brain-Machine Interfaces (BMIs), focusing on specific concerns related to safety, autonomy and responsibility, and technical aspects of these technologies. The development of BMIs in connection to AI algorithms enable more accurate decoding of neural signals, facilitating applications in neuroprosthetics, cognitive enhancement, and even direct brain-to-brain communication. Even though this is positive and can help numerous patients with specific diseases, several ethical concerns should be raised: (i) BMIs can pose a direct risk of harm, particularly for invasive methods that require surgical interventions related with implants, since there is potential for complications including infections surrounding the tissue or acute trauma to the brain (Wolpaw and Wolpaw, 2012); for long-term implants, the affected neural tissue may also develop glial scarring, which can surround the implant and obstruct the BMI function (Borton et al., 2013); (ii) issue regarding autonomy and responsibility could also arise from BMIs: is an action primarily or exclusively generated by a

¹ Research Fellow, Mind Language and Action Group, Institute of Philosophy, University of Porto, Portugal. Honorary Professor, Faculty of Medicine, University Andrés Bello, Chile.

device still attributable to a human? If a BMI device significantly influences an individual's decision-making process, this could potentially undermine autonomy; for example, studies found that the illusion of agency is a reality, where BMI patients inaccurately claim to be the agent of action when that was not the case (Vlek et al., 2014); finally, (iii) there are several technical issues that needs to be considered from a normative point of view: BMIs rely on accurately capturing neural signals: noise, signal degradation, and interference can affect the quality and reliability of the signals, leading to inaccurate or unreliable control of devices; over time, neural signals can change due to various factors, such as electrode drift or biological changes, and this can lead to a decrease in the accuracy and stability of the BMI's performance; moreover, interference between neighbouring electrodes can lead to crosstalk, where the signals from one electrode affect the measurements from others. Neural crosstalk can complicate accurate signal interpretation (Clausen, 2011). The chapter concludes by addressing the need for transparent AI algorithms in BMIs, advocating for ethical oversight to ensure that these innovations prioritize sufficient ethical standards, while balancing the benefits and risks of neurotechnology.

Keywords: Brain-Machine Interface, Artificial Intelligence, Ethics, Autonomy, Privacy.

1. Introduction

1.1 Background context

The fusion of Brain-Machine Interfaces (BMIs) and Artificial Intelligence (AI) marks, for some, a key moment in the development of neurotechnology broadly conceived: these systems are no longer mere lab experiments, but are here and now entering clinical trials and towards human-based applications with an interesting cognitive and

therapeutical potential (van Stuijvenberg et al., 2024a). Fan et al. (2023) reported stable decoding accuracy of 93.8% over 403 days – more than one year – in a real-time BCI system (intracortical handwriting), making use of continual online recalibration framework, which showed a relevant performance in a human participant, underscoring mutually their technical feasibility and clinical capability.

Nevertheless, this technical progress carries ethical complexities: the potential to be able to decode human intentions or emotions with these kinds of interfaces make us confronted with several philosophical questions about personal identity, privacy, and autonomy: for example, Yuste and Goering (2017) argue that these technologies risk undermining fundamental human values, demanding frameworks that can protect privacy, identity, agency, and equality. In the same line of thought, van Stuijvenberg et al. (2024b) underline that user control ought to be highly relevant from an ethical point of view: users must not be instrumentalized, but their agency ought to be explicit, controllable and clear for both themselves and other people.

Moreover, national-level ethical efforts have been rising to lead what can be called “responsible” BMI deployment: for example, China’s 2024 Guidelines for Research Ethics in Brain-Computer Interface (NRECChina, 2024) advocate sustainable governance via alignment with already established medical protocols (e.g., the Declaration of Helsinki), while UNESCO and European bodies are concurrently developing similar frameworks (UNESCO, 2024). These initiatives indicate a growing concern focused on the idea that AI-inspired BMI technologies in general cannot overtake the ethical and legal frameworks meant to steward it. The aim of this chapter will be to

explore how, in different contexts, AI-driven BMIs may be already overtaking some fundamental ethical concerns.

1.2. Definitions and Scope

Before starting the ethical analysis, let's make it clear what kind of technologies this chapter will be focusing on. First, Brain–Machine Interfaces (BMIs, also referred in the literature as BCIs) are systems that can record activity from the central nervous system (whether via implanted, invasive sensors, or non-invasive scalp electrodes) and translate these signals in (nearly) real-time into functional outputs that replace, restore, or augment natural brain capabilities (Wolpaw et al., 2002; Wolpaw & Wolpaw, 2012). This broad classification is recognized by the BCI Society and reproduces a consensus in neurotechnology scholarship: BMIs enable users to communicate or control devices directly through neural signals (BCI Society, 2024).

Secondly, and within the BMI context, Artificial Intelligence (AI) refers to advanced machine-learning models and deep neural networks that are used to process and identify intricate neural signal patterns: these AI models are able to decode user intentions into specific actions by detecting and deciphering motional or cognitive commands from noisy, high-dimensional neural data (Barredo Arrieta et al., 2019).

This chapter will, therefore, interrogate the overlapping ethical dimensions of the crossing of these technologies: how AI-enhanced BMIs challenge our traditional concepts of responsibility, autonomy, and safety. Some of the questions we will touch on will be the following: Can users of AI-BMIs be (truly) autonomous in cases where their actions emerge from technological mediation? Also, what are the boundaries of responsibility in cases when an AI model triggers

particular bodily action? And how do we guarantee that these systems are safe, reliable, and transparent for both the users and experts?

To answer to these queries, we will advance with three main goals that will offer the kind of relevant normative analysis that we consider to be relevant to understand the ethical reflection of AI-BMIs: first, we will consider both invasive (e.g. intracortical implants for motor prostheses) and non-invasive BMI (e.g. scalp-based neurofeedback in locked-in syndrome patients) so we can contextualize, in the clinical practice, how AI can be deployed and used to achieve different concrete outcomes. Moreover, we will try to identify specific ethical tensions that can be raised by these BMI-AIs tools, discussing three ethical concerns:

- (a) how the autonomy and agency of patients can be disrupted by the mediation of these technological developments (Vlek et al., 2014);
- (b) how the physical safety of patients – particularly in the long run –, can be problematic due to the high risk of infection, inflammation and surgical complications that are hard to predict (Ryu and Shenoy, 2009);
- (c) how the technological reliability can be guaranteed when faced with possibilities of “crosstalk”, signal or biological degradation or issues of lack of transparency of many AI systems (Barrese et al., 2013).

Finally, we present a normative framework that will help us to support our ethical analysis, grounded in the idea that the technology used in AI-BMIs scenarios ought to be transparent for users, medical experts and everyone involved in the development and use of the tool. We will also propose an ethical oversight structure based on the idea of expert ethical boards, one that can emerge from the field of international Neuroethics (van Stuijvenberg et al., 2024) similar to

what happened with Ethics Committee for hospitals and universities in the previous century.

Following this tripartite structure (mapping, evaluation, and normative framing) we will be able to create a relevant narrative that will take both empirical descriptions to moral analyses in a constant dialect in order to achieve some appropriate practical recommendations that, hopefully, may be considered by those developing these kinds of technological developments.

2. Technological Integration: AI and BMIs

2.1 Brief History and Current State of BMI Development

The first Brain–Machine Interface (BMI) development traces back to some of the foundational experiments in bioelectricity, by Hans Berger: in 1924, he produced the first human electroencephalogram (EEG), identifying both α - and β -wave patterns and signaling the dawn of neural signal measurement (Berger, 1929). Developing this and other methods, researchers were able to interfacing directly with brain activity: Delgado's invasive implants demonstrated neural stimulation in both animals and humans, and Vidal coined the term “brain–computer interface” in 1973 (demonstrating single-trial EEG control of cursors (Delgado, 1969; Vidal, 1973).

Based on these breakthroughs, Kennedy and Bakay were able to use the first human intracortical BMI implant in 1998 in a patient suffering of a locked-in syndrome (Kennedy & Bakay, 1998). Seven years after, a tetraplegic patient was able to use its own neural signals to control a robotic arm, implanting 96 electrodes (Hochberg et al., 2006). The same kind of achievement was made through non-invasive methods: for example, the SMR-based cursor navigation system developed in

the early 1990 by Wolpaw in a successful case of BMI without any invasive component (Wolpaw et al., 2002).

In the last decade, the diversity and scale of BMIs has been raising with many different projects and applications: for example, DARPA-funded initiatives and private ventures (e.g., Neuralink, Kernel, Facebook) are now exploring ways to increase the viability and bidirectionality of both invasive and non-invasive methods of BMIs (DARPA, 2019): invasive methods are based on intracortical electrodes that are implant in the brain with high signal and spatial resolution, while the non-invasive methods (such as EEG, ECoG or NIRS) can provide less risk of harm and dependence in the long-run (Gosh, 2023).

Now, the goal of integrating Artificial Intelligence into BMIs can be conceptually confusing since it broads the functional boundaries of what is a traditional technology

The integration of Artificial Intelligence (AI) into Brain–Machine Interface (BMI) systems has redefined the conceptual and functional boundaries of neurotechnology: at first, since more rudimentary tasks were needed (e.g. cursor control, binary selection task, etc.), linear decoding models such as the Wiener and Kalman filters were more than enough (they assume a sort of stable relationships between the task (i.e. motor output) and the intention (i.e. neural inputs) (Kim et al. 2008). However, for more complex tasks, these models are very problematic when applied to high-dimensional neural data that are non-stationary by definition, which can be found in multi-task or long-term contexts.

Because of this, data-driven models using machine and deep learning had to be replace these deterministic models: the traditional methods (e.g. Linear Discriminant Analysis or the Support Vector Machines),

even if they were quite effective in processing signals from noise than the previous methods, cannot achieve the kind of decoding accuracy, specificity (for different users) and generalizability that these new models can (Glaser et al., 2017).

The raising and applications of deep neural networks (e.g. Convolutional Neural Networks and Long Short-Term Memory) have shown to be more efficient in processing the complex spatial-temporal features that arise from different neural signals of complex task: these new models can identify the distribution of complete neural patterns and can, for example, decode speech from cortical activity, raise the ability of motor control or help users to reconstruct visual imagery (Hu et al., 2023).

This new development, the AI-BMIs, can allow for co-adaptive decoding methods that can adapt way more easily to the intra-subject variability that will naturally occur over time to a patient: by using online and reinforcement learning methods, this can help to face what is known as “neural drift”, signal degradation and, more importantly, the electrode reconfiguration. If we consider patients with chronic implanted systems, this can allow achieving a long-term stability and adaptability that is very useful from a clinical point of view (both for the efficiency of the tool, but also for the well-being of the patient) (Orsborn, et al. 2012).

Some recent applications of this kind of AI-BMIs technology can be mentioned: for example, Nvidia partnered with Synchron by developing specific edge-AI modules (Holoscan platform by Nvidia) which enable low-latency decoding of wireless stentrode BMI (a less invasive method that implants a stent-like device not in the brain, but in a blood vessel near the brain) (Synchron, 2025). This improves both the technical difficulties present in more traditional methods, but also

raises the commercial viability of these technologies, which can allow, in a near future, to diversify the applications and uses.

This novel integration of Artificial Intelligence into Brain–Machine Interface technology, even if it can bring many positive outcomes from a clinical point of view, raises many ethical and social queries that also need to go hand-a-hand with the technical development. For example, it is not clear at all if data privacy or signal interpretability can be fully guaranteed: the application of AI methods can develop specific algorithmic bias that were not found in more traditional versions of BMIs. As it is well-known, neural data are immensely personal and, therefore, sensitive from a privacy perspective: they can carry, at least in theory, the potential to reveal more than just the expected clinical outcome (e.g. a motor intention) but also, and perhaps even more, emotional, cognitive and thinking states (Zhou et al., 2022).

If we consider the problem of interpretability of many AI models that can be applied to BMIs, then we can have similar “AI-BMI” black box problem that we see currently happening with some deep learning architectures or generative transformers: since they are quite obscure – from an epistemic point of view –, this may raises consequences for the kind of trust users and experts can have in this technology, but also the in the kind of clinical validation that these methods should achieve to be widely used (Varshney and Alemzadeh, 2017).

Other ethical considerations that should be relevant to consider are the potential to – and again, at least in theory – allow cognitive or behavioral influence on those using the technology in question: this raises obvious concerns regarding traditional concepts of medical bioethics such as autonomy, informed consent or the integrity of personal identity (Yuste et al., 2017). To avoid missing the “ethical train”, we will propose that specific multidisciplinary regulatory frameworks (one that considers the collaboration of different

expertise) ought to be considered before the technology is widely available and used in clinical or personal contexts.

Less direct but also highly relevant are the socio-economic inequalities that can be worsened even more: on one hand, this technology is very costly and, therefore, it may be only accessible to those with economic power and, by consequence, inaccessible to those that cannot acquire these clinical tools (Rainey et al., 2020); on the other hand, the potential misused in other contexts – e.g. military, surveillance –, demands for careful oversight on how this technological development will be developed and deployed (Ienca, Jotterand, & Elger, 2018).

Next, we will analyze some of the key applications of this neuro-technology in clinical and other contexts to provide empirical contexts relevant for the ethical analysis.

2.2. Key Applications in AI-BMIs and a Case Study

As we noticed, Artificial Intelligence-driven Brain–Machine Interfaces (AI-BMIs) are starting to transform many different applications in several domains (e.g. cognitive, communicative or clinical). These new kinds of neuroprosthetics can directly translate neural signals into a wide range of commands which can be applied to control wheelchairs, prosthetic limbs or even full exoskeletons, restoring patients that were affected, for different reasons, by strokes, limb loss or paralysis (Wolpaw and Wolpaw, 2012).

The applications of AI can specifically enhance both the speed and accuracy of the BMI, which provides, from a user point of view, more natural and intuitive control. Moreover, AI models can help to increase the level of adaptability for both calibration and error correction of neural signals, which allows for a higher user comfort and durability in the long run (Horváth, Kovács & Benkő, 2024).

Psychologically, these developments also can affect the reintegration of patients using the technology, since they can be inserted in social and occupational activities that, before this use, could not happen.

If we consider cognitive enhancement, AI-BMIs can potentially augment different aspects of the human mind: with their capacity to real-time detection and constant modulation of neural patterns connected to working memory, executive function or attention, this can open the door to different kind of enhancement that can raise normative considerations since it can allow humans to perform beyond normal levels of performance (Tsai et al. 2025). The same technology can also be applied to increase the capacity of treatments in neuropsychiatric disorders: through AI-drive neurofeedback systems, patients with unbalanced brain functions (e.g. anxiety or ADHD) can receive closed-loop interventions (which are tailored to each individual's neural networking) with promising results (Sitaram et al., 2017). For healthy individuals, several non-invasive techniques (e.g. transcranial direct current stimulation or transcranial magnetic stimulation), merged with AI models, are currently being explored and tested to enhance both the learning and cognitive performance (Zhang et al. 2020).

A particularly innovative application of AI-BMI can actually be considered not a BMI (at least in the traditional sense): brain-to-brain communication (B2BC) is a new way to set up different mental agents (e.g. two persons, two monkeys, etc) and share relevant information in both to achieve a specific and mutual task: some experimental studies have been already performed using non-invasive EEG and transcranial stimulation, with the support of AI models to decode and encode neural signals to share commands, abstract information of images (Rao et al., 2014). Of course, this new technology, being a further development of more traditional BMIs, are accompanied by

the same ethical problems identified before, such as concerns for privacy, autonomy and the influence on the self of those using the technologies. However, it also raises new,² original ethical problems, such as the problem of responsibility in collective decision-making (that is, when two people are interacting using a B2BC, who ought to be responsible by the outcomes of that use?) or in the raise of new kind of social interactions never seen before (Yoo et al., 2023).

An incredible case study of BMI influenced by AI is the pioneering research developed by Miguel Nicolelis and his collaborators. In one of the most emblematic moments for humanity, a paraplegic patient named Juliano Pinto was implanted with a BMI system with the goal of performing the first official kick ball of the World Cup 2014 in Brazil by controlling a full exoskeleton (BBC, 2014).

The BMI used was an invasive one, where many microelectrodes were implanted within Pinto's primary motor cortex focused on the leg movement: this allowed for a recording (in high-resolution) of action and field potentials, providing a complex map of data that recreated the motor intentions of the user. But this was not a "simple" BMI, but an AI-based technology that was able to decode and translate the complex neural patterns into concrete actions directed to the robotic exoskeleton: machine learning methods were used (from linear and non-linear to more adaptative techniques) in order to capture, on one hand, the dynamic nature of neural activity and, on the other hand, the non-stationary brain patterns, allowing the BMI to maintain relevant control of the technology even with high neural changeability

² B2BC also raises new technical challenges such as improving information bandwidth, reliability, and user training protocols which have indirect ethical concerns.

and signal drift (Nicolelis and Lebedev, 2009; Nicolelis and Lebedev, 2014).

To understand the complexity of the application of this kind of technological tool, it is interesting to describe some of the incredible process that Pinto's add to endure in order to be able to control, with its own mind, a complex exoskeleton. After the surgical implantation of the electrodes, the patient had to record his neural activity in both real and imagined motor movements during several weeks: the BMI design allowed to incorporated co-adaptative learning paradigms where decoding algorithms were used to optimize the performance. This bidirectional arrow is highly relevant to achieve the kind of natural fluidity to perform a soccer kick: of course, the complexity of the robotic exoskeleton was also important since it recreated the appropriate biomechanics for that particular movement (Colucci et al., 2022).

This technical achievement shows the potential of AI-BMIs for clinical use but also showed how invasive methods can be quite critical due to their surgical risks and long-term biocompatibility (in the case study used, it is not clear that Pinto will be able to access the exoskeleton for other more important tasks than kicking a soccer ball). Also, it showed how the perception of the nature of human disability can change with the use of technology broadly considered.

With this brief overview of the AI-BMI current paradigm, we will now focus on the ethical challenges that are raised by these technologies. We will provide a framework to evaluate, concretely, its normative limits and standards.

3. Ethical Concerns in AI-driven BMIs

3.1 Invasive vs. Non-Invasive BMI Methods (AI-Driven)

As previously mentioned, AI-BMIs can be divided into two main modalities: invasive and non-invasive methods. Each of them represents different clinical and technical details and, more importantly, specific ethical considerations that ought to be made.

Invasive BMIs usually involve those applications that require a direct implantation, in the cortical surface or the parenchyma of the brain, of electors that enable to acquire and replicate very precise neural signals to control external devices: the signals have to be decoded and interpreted by AI algorithms which excel at identify temporal and spatial patterns from the brain activity (Jiang et al., 2023).

Differently, non-invasive BMIs mainly depend on external sensors such as functional near-infrared spectroscopy (fNIR), magnetoencephalography (MEG) or, more famously, electroencephalography (EEG), which enables to monitoring of the neural activity not via the brain, but the scalp or skull. The idea here is to avoid the physical risks that any brain surgery has (more soon): but to avoid this, something else is lost, namely, a lower spatial resolution in conjunction to higher levels of “noise”, which influence the efficiency of the technology. Here as well AI-driven methods have been relevant to enhance these two problems: by applying specific models, non-invasive methods can increase the classification accuracy, feature extraction and improve the signal decoding (Zhao et al., 2021).

Some of these techniques are RNNs (recurrent neural networks), SVM (support vector machines) or CNNs (convolutional neural networks), which can increase the interpretation of the non-invasive signals and, therefore, rise the overall accuracy of the BMI in use.

If we take into consideration the clinical perspective, BMIs have been quite useful regarding the performance of fine motor control or multi-degree-of-freedom commands (e.g. speech decoding, robotic arm/leg movement, etc.) (Hu et al., 2023). Invasive BMIs are more interesting in this task, since they can take into account the neural plasticity and the signal non-stationarity, which allows to avoid changes of performance over time. Nevertheless, as we will see below, invasive BMIs have high surgical risks and biocompatibility limits that should be ethically considered, especially if we take into consideration the widespread application of this technology for different populations. This is where non-invasive BMIs can have an important role: if we look at safety and ease of use, this kind of technological application seems better to monitoring cognitive states, to help locked-in-patients communicate or to provide neurofeedback therapies (Krigolson et al., 2020).

From a normative perspective, the decision to embrace invasive vs non-invasive methods raise pertinent questions about patient autonomy and well-being, trade-offs benefit between harm and efficiency, and the kind of informed consent that should be based on known factors that can be predicted over time. If it is true that invasive BMIs offer high control capabilities, the long-term and neural impact of implementing electrodes in the brain of patients ought to be clearly indicated; on the other side, non-invasive BMIs are safer (when compared to the invasive ones), but they face high limitations regarding the functionality and reliability of the technology, which can affect in some cases the autonomy of users (e.g. in cases where the technology does not perform what the user is intending) or the dignity of users (e.g. when a technology is been used but does not bring an practical relevant advantage).

Next, let's detail some of the risks that can occur in Invasive BMIs.

3.2. Risks of Surgical Intervention in BMIs

As we saw before, invasive BMIs are interesting from the efficiency point, but it introduces many different medical risks that can have both clinical and ethical considerations. Let us consider some of the stages required to implement a common BMI system, such as intracortical implants.

First of all, it should be noticed that these kinds of procedures are very expensive from a resource perspective: they require a coordination between engineering, neuroscientific and clinical areas that develop, throughout a timeline that can extend from several months to years (depending of the complexity and the degree of motor control intended), the entire system.³

The process starts with a pre-surgical phase that is non-invasive and is focused on determining the cortical targets of the patients: this involves high-resolution brain imaging such as fMRI, MRI or CT crossed with neuropsychological data (typically from the primary cortex motor when focused on motor BMI) (Flesher et al., 2021). In this stage, informed consent measures are taking place even if the long-term implantation are highly correlated with unknown risks, which creates a very delicate ethical situation in these cases.

The next phase is the actual surgical implantation of the electrodes: this happens in full anesthesia which can take up to 6 hours to conclude, depending of the kind of device chosen (e.g. ECoG grids, Stentrodes, etc). The post-surgical recovery can vary in time depending of the age and fitness of the patient, but it typically takes 2

³ Neurosurgeons, biomedical engineers, neuroethicists, rehabilitation specialists, and AI developers are all involved in this process. Clinical trials often report resource use exceeding \$100,000 USD per participant per year (Daly and Wolpaw, 2008).

weeks; after, the first signal quality testing happens: for example, Neuralink use 96-channel microelectrode to interface with external devices that decode the signals to achieve, over the course of many weeks, the kind of calibration needed. This takes daily sessions of training between 1-3 hours where a specific task is made (e.g. imaging an image and record the neural activity) (Orsborn et al., 2012).

In this stage, machine learning models (e.g. SVMs or recurrent neural networks) are being trained to identify neural patterns and cross those with the motor intentions. For that, it is needed at least 40 hours of recorded data that are constantly refined to achieve the best results possible (Collinger et al., 2013). Studies on full functional control of robotic limbs showed the need of intensive training between 3 to 6 months, depending of both the cortical baseline plasticity of the patient and the kind of task performed. During this time, many things can happen, from signal drifting to electrode degradation, which require a laborious and demanding constant technical work.

If we go back to the initial story of the 2015 World Cup kick, Juliano Pinto had to train for months not only the neural phase, but also the linking with the effector system (in this case, the exoskeleton), calibrating it to match the natural velocity, biomechanics and trajectories of the patient.

Looking at the different stages, we can see how much risk is embedded in this kind of procedures: infections and tissue damages can cause severe complications and neurological discrepancies, at least in the long-term. Since the electrodes have to breach the protection of the skull and meninges, they can easily become a source of microbial infection, which can create all kind of clinical problems (e.g. meningitis, sepsis, abscess, etc.) which may need aggressive treatments or even, in worse cases, the removal of the implant, exposing the patient to even more risks while losing all the potential

therapeutic benefits (Wolpaw and Wolpaw, 2012). It should also be clear that the insertion of electrodes disrupts neural tissues which can cause injuries to several neurons (and glial cells) due to the hemorrhage, causing inflammatory responses that can lead to further neuronal loss.

The motion between the actual brain tissue of the patients and the rigid electrode during the normal use of the device, since they are biologically different, can cause many tissues trauma and chronic tissue stress, which explains why there is a signal degradation over time. Moreover, the glial scarring can also cause negative effects on surrounding brain structures, which can lead to more neural degeneration and can affect the overall efficiency of the technology, which require a biocompatibility between the implant and the brain tissue that is hard to achieve currently (Otte, et al., 2022).

If we take all these empirical details into consideration, we can easily see how the surgical implantation causes some ethical reserves in the risk-benefit calculus of invasive BMIs: on one hand, the principle of non-maleficence requires diminishing harm and therapeutic gain if harm is done; but it is not clear that these can be guaranteed by the multidisciplinary team, even if informed consents may include an exhaustive list of the risks. We also know that many of these patients are in a fragile epistemic state due to their unhealthy situation and, therefore, are more likely to accept risks that they would never accept if healthy.

Therefore, we can claim that both infection and tissue damages, in the long run, are causally associated with this method and, therefore, we have a moral need to consider adequate responses to these long-term negative effects: addressing them will require better science, that is, will require design improvement, developing more flexible and

biocompatible electrodes that can avoid infections and damage in the implantation area (Kozai et al., 2015).

3.3. Autonomy and Responsibility in BMIs

When we think in traditional concepts applied to healthy individuals, the notion of autonomy and responsibility are more or less easily applied if you ignore the metaphysical debates on these two tricky concepts. But when we apply those standards definitions in context of BMIs, things may get philosophically obscure since these concepts are related to conscious intentionality and control over the movements of our bodies. Since BMIs can act exactly on that (i.e. conscious and movement), they introduce new modes of interaction and, by consequence, of relevant impact in the world. Therefore, we can ask: what can count as a voluntary act of the user when using an interface that modulates the interaction between is conscious intentions and the specific execution of a movement? (Mele, 2009)

Firstly, if we take into consideration the fact that some of the neural signals that control a movement may not fully reflect a conscious intention (in the full sense), particularly if you take into consideration involuntary fluctuations in different areas of the brain that can be wrongly decoded, this can result in unintended performances from the user (Wolpaw and Wolpaw, 2012). Second, even if training and mental strategies can help to calibrate the control of the device, latency and errors can always occur, transforming control in a probabilistic system rather than a deterministic one. But if it is probabilistic, then a direct causal relationship between actions and intentions cannot be made and, therefore, this undermines the traditional concepts of responsibility we use (Hochberg et al., 2012).

Finally, when we use semi-autonomous (or even fully autonomous) AI models to decode or optimize signal performance, then this

becomes even more problematic: we are dealing with a kind of shared agency that is not ethically considered in our legal and moral codes. Consequently, there is an urgent need to reconceptualize our concepts of autonomy and responsibility to one that takes into consideration a distributed perspective that is more relational than binary, that can provide a clear spectrum of fully intentional to partial or unintended actions (Gouveia, 2019).

This will demand an upgrade from intentional ethics to a kind of non-intentional actions, where human agents are no longer the only players in town: hybrid agent systems are arising and we need to develop relevant ethical frameworks to deal with them. This is especially relevant when we have some relevant empirical research that show that BMIs create, often, an illusion of agency, when users consider that an action is their initiative (Vlek et al., 2014). This disruption of agency challenges how users using BMIs perceive their own actions and the authenticity of their acting: there is a clear dissociation between what is the subjective experience of an intention and the objective control of that intention. If you take, for example, into consideration the delay of signal processing in some contexts, this temporal disruption can clearly affect the agency of the user, even if they perceive it as a legitimate act (Haggard, 2017).

These challenges require a level of care and attention by BMI experts in the training phase that may ensure that patients are aware of the kind of limits and alienation phenomena that may occur under the use of a BMI, in order to respect its autonomy but also its well-being in the long run.

3.4. Technical Challenges of BMIs

To conclude, we review some of the most fundamental technical challenges of BMIs that should be taken into accounting when thinking the morality of these technological applications.

First, as it was mentioned before, there are challenges regarding the signal reliability that can be achieved with AI-powered BMIs: even if it is true that this new version works better than traditional BMIs, the quality and stability of brain signals over time affects directly the performance which, in the long run, poses a significant challenge. The signal degradation can also occur with micromovements of the implant that can cause tissue reactions that reduce the functioning ratios (signal-to-noise) which affect, by consequence, the mental representation relevant for a specific task (Campbell and Wu, 2018). Non-invasive methods solve some of these problems, but they offer a low spatial resolution and can be easily interfered by external sources, which affects their reliability in the short time.

We can, however, claim that what is needed is actually better AI tools that can recalibrate, over time, the signal fluctuations, enabling a better performance even if degradation occur. Nevertheless, the use of these kinds of AI models introduces their own philosophical problems: specifically, many of these models have a black-box structure, which raises transparency issues that challenge the autonomy of patients in general (Gouveia, 2025). They can lead to algorithmic bias or difficulties in detecting failures which cannot be identified and, by consequence, corrected. Explainable AI frameworks can be an alternative, but these methods are still being explored in the present (Gouveia & Malík, 2024)

A second issue that is raised is the possibility of crosstalk or interference of the functioning of the implanted electrode. The typical

invasive methods rely on multi-electrode arrays which can affect the individual performance of each electrode by unintentionally overlap or mix different signals of the adjacent electrodes, degrading the relevant neural signals for a specific task (Dunlap et al., 2020). There are several factors that can cause this phenomenon: resistance incongruity, neural plasticity, electrical noise and the physical proximity of each electrode, all of those affecting the quality of the signals recorded, which can lead to an inaccurate outcome for the user (Nicolelis and Lebedev, 2009).

To solve this issue, experts are trying to focus on improving the hardware design of the electrodes with enhancing the techniques for signal processing: this may involve, on one hand, improving the materiality⁴ of which the electrodes are made of and their spatial distribution in the brain and, on the other hand, develop deep learning techniques to blind each electrode's activity and avoid the overlapping of signals to improve the decode precision (Liu et al., 2024)

From an ethical point of view, the possibility of crosstalk raises higher demands that we currently have in the field, since they affect both the efficiency but also the safety of users using an AI-BMI, which can lead to loss of trust in the technology (and psychological frustration or harm) or, in worse scenarios, actual physical harm. Therefore, we can claim that ensuring the best signal performance possible is not only a technical goal but actually, a normative standard that should be in place in every single team that develop BMIs applications.

⁴ Flexible, ultrathin, and bioactive-coated can reduce scar formation and extend recording stability.

4. Conclusion: Some Ethical Tools

We saw in this chapter how AI-driven Brain-Machine-Interfaces raise many interesting clinical applications but also some ethical queries that we ought to consider. To deal with them, a couple of traditional ethical perspectives should be in place.

First, there is a need to take seriously both ethical review boards or committees and regulatory frameworks adapted to the specific details of this technology: given the impactful nature of BMIs in the mental life of users and the potential risks associated (e.g. mental privacy, disruption of personal identity, impact on the control and agency, etc) (Ienca and Andorno, 2017). Therefore, we need regulations that focus not only in the physical/clinical impact, but also in the mental “freedom” and respect of personal autonomy of users.

Ethics committees should also have a role here in applying these regulatory frameworks: they should be multidisciplinary by nature (in order to incorporate a holistic perspective of a human being), and they should act not only in the approval phase, but constantly throughout the different stages of the development and use of the technology. At the same time, there is an urgent need for what can be known as “international” harmonization of these different regulation and committee roles to be sure to avoid at all cost “ethics dumping” (Floridi et al., 2018). Finally, create an ethical culture and environment within the institutions and industries that develop AI-BMIs should be mandatory, where ethical reflection is seen as an essential feature of the work produced instead of a mere secondary aspect.

In sum, the amazing clinical potential of AI-BMIs necessitates a sustained ethical approach that can soundly informed policy-making and regulation while holistically approach neurotechnologies having

in human relevant societal values that ought to be always present and never forgotten.

References

- BBC News (2014), 'Paraplegic Man Controls Robot to Kick Off Brazil World Cup', 12 June, Available at: <https://www.bbc.com/news/technology-27714214> (Accessed: 13 June 2025).
- Borton, D.A., Yin, M., Aceros, J. & Nurmikko, A. (2013) "An implantable wireless neural interface for recording cortical circuit dynamics in moving primates", *Journal of Neural Engineering*, 10(2): 026010.
- Campbell, A., & Wu, C. (2018) "Chronically Implanted Intracranial Electrodes: Tissue Reaction and Electrical Changes", *Micromachines*, 9(9): 430.
- Clausen, J. (2011) "Conceptual and ethical issues with brain-hardware interfaces", *Current Opinion in Psychiatry*, 24(6): 495–501.
- Collinger, J., Wodlinger, B., Downey, J., Wang, W., Tyler-Kabara, E., Weber, D., McMorland, A., Velliste, M., Boninger, M. & Schwartz, A. (2013) "High-performance neuroprosthetic control by an individual with tetraplegia", *The Lancet*, 381(9866): 557–564.
- Colucci, A., Vermehren, M., Cavallo, A., Angerhöfer, C., Peekhaus, N., Zollo, L., Kim, W., Paik, N., & Soekadar, S. (2022) "Brain-Computer Interface-Controlled Exoskeletons in Clinical Neurorehabilitation: Ready or Not?", *Neurorehabilitation and neural repair*, 36(12): 747–756.
- Daly, J. and Wolpaw, J. (2008) "Brain-computer interfaces in neurological rehabilitation", *The Lancet Neurology*, 7(11): 1032–1043.
- Dunlap, C., Colachis, S., Meyers, E., Bockbrader, M. and Friedenberg, D. (2020) "Classifying Intracortical Brain-Machine Interface Signal

- Disruptions Based on System Performance and Applicable Compensatory Strategies: A Review", *Frontiers in Neurorobotics*, 14:558987. doi: 10.3389/fnbot.2020.558987.
- Fan, C., Hahn, N., Kamdar, F., Avansino, D., Wilson, G. H., Hochberg, L., Shenoy, K. V., Henderson, J. M., & Willett, F. R. (2023). Plug-and-Play Stability for Intracortical Brain-Computer Interfaces: A One-Year Demonstration of Seamless Brain-to-Text Communication. *Advances in neural information processing systems*, 36, 42258–42270.
- Flesher, S., Collinger, J., Foldes, S., Weiss, J., Downey, J., Tyler-Kabara, E., Bensmaia, S., Schwartz, A., Boninger, M. and Gaunt, R. (2021) "Intracortical microstimulation of human somatosensory cortex", *Science Translational Medicine*, 13(598): eabe0662.
- Floridi, L., et al. (2018) "AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations", *Minds and Machines*, 28(4): 689–707.
- Gouveia, S. (2019) "A Defense of the Principle of Distributed Responsibility in Artificial Intelligence", *Revue Roumaine de Philosophie*, 63(2): 235–248.
- Gouveia, S. (2025) "The Ethics of Artificial Intelligence in Medicine: Preliminary Remarks", *Global Philosophy*, 35, 4.
<https://doi.org/10.1007/s10516-024-09737-y>.
- Gouveia, S. and Malík, J. (2024) "Crossing the Trust Gap in Medical AI: Building an Abductive Bridge for xAI", *Philosophy and Technology*, 37, 105. <https://doi.org/10.1007/s13347-024-00790-4>.
- Haggard, P. (2017) "Sense of agency in the human brain"; *Nature Reviews Neuroscience*, 18(4): 196–207.
- Hochberg, L., Bacher, D., Jarosiewicz, B., Masse, N., Simeral, J., Vogel, J., Haddadin, S., Liu, J., Cash, S., van der Smagt, P., and Donoghue, J. (2012) "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm", *Nature*, 485(7398): 372–375.

- Horváth, Z., Kovács, A., & Benkő, Z. (2024) "Enhancing motor imagery classification in brain-computer interfaces using deep learning and continuous wavelet transform", *Applied Sciences*, 14(19): 8828.
- Hu, Y., Zhang, Y., Zhang, S., Wang, X. and Sun, H. (2023) "EEG-based emotion recognition using hybrid CNN-LSTM with self-attention", *Biomedical Signal Processing and Control*, 82, p.104589.
- Ienca, M., Andorno, R. (2017) "Towards new human rights in the age of neuroscience and neurotechnology", *Life Sci Soc Policy*, 13, 5. <https://doi.org/10.1186/s40504-017-0050-1>.
- Ienca, M., Jotterand, F., & Elger, B. S. (2018) "From Healthcare to Warfare and Reverse: How Should We Regulate Dual-Use Neurotechnology?", *Neuron*, 97(2): 269–274.
- Kozai, T., Jaquins-Gerstl, A., Vazquez, A., Michael, A. and Cui, X. (2015) "Brain tissue responses to neural implants impact signal sensitivity and intervention strategies", *Annual Review of Biomedical Engineering*, 17: 45–72.
- Koralek, A., Jin, X., Long, J., Costa, R., & Carmena, J. (2012) "Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills", *Nature*, 483(7389): 331–335.
- Krigolson, O., Williams, C., Norton, A., Hassall, C., and Colino, F. (2020) "Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research", *Frontiers in Neuroscience*, 14: 1–14.
- Liu, X., Gong, Y., Jiang, Z., Stevens, T. and Li, W. (2024) "Flexible high-density microelectrode arrays for closed-loop brain-machine interfaces: a review", *Frontiers in Neuroscience*, 18:1348434. doi: 10.3389/fnins.2024.1348434.
- Mele, A. (2009) *Effective Intentions: The Power of Conscious Will*, New York: Oxford University Press.

- Nicolelis, M. and Lebedev, M. (2009) "Principles of neural ensemble physiology underlying the operation of brain-machine interfaces", *Nature Reviews Neuroscience*, 10(7): 530–540.
- Nicolelis, M. and Lebedev, M. (2014) "Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation", *Physiological Reviews*, 97(2): 767–837.
- Orsborn, A.L., Dangi, S., Moorman, H.G. and Carmena, J.M. (2012) "Closed-loop decoder adaptation on intermediate time scales facilitates rapid BMI performance improvements independent of decoder initialization conditions", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(4): 468–477.
- Otte, E., Vlachos, A. & Asplund, M. (2022) "Engineering strategies towards overcoming bleeding and glial scar formation around neural probes", *Cell and Tissue Research*, 387(3): 461–477.
- Rainey, S., Martin, S., Christen, A., Mégevand, P., & Fournieret, E. (2020) "Brain Recording, Mind-Reading, and Neurotechnology: Ethical Issues from Consumer Devices to Brain-Based Speech Decoding", *Science and engineering ethics*, 26(4): 2295–2311.
- Rao, R., Stocco, A., Bryan, M., Sarma, D., Youngquist, T., Wu, J., and Prat, C. (2014), "A direct brain-to-brain interface in humans", *PLOS ONE*, 9(11), e111332.
- Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Rana, M., Oblak, E., Birbaumer, N. & Sulzer, J. (2017) "Closed-loop brain training: the science of neurofeedback", *Nature Reviews Neuroscience*, 18 (2): 86–100.
- Tsai, P, Akpan, A., Tang, K. and Lakany, H. (2025) "Brain computer interfaces for cognitive enhancement in older people - challenges and applications: a systematic review", *BMC Geriatrics*, 25, 36.<https://doi.org/10.1186/s12877-025-05676-4>

- van Stuijvenberg O.C., Samlal DPS, Vansteensel MJ, Broekman MLD and Jongsma KR (2024b) "The ethical significance of user-control in AI-driven speech-BCIs: a narrative review", *Frontiers in Human Neuroscience*, 18: 1420334. doi: 10.3389/fnhum.2024.1420334.
- Vlek R, Van Acken J, Beursken E, Roijendijk L, Haselager P. (2014) "BCI and a user's judgment of agency". In: Grubler, G and Hildt, E. (eds.) *Brain-computer-interfaces in their ethical, social and cultural contexts*, Dordrecht: Springer, pp. 193–202.
- Yoo, S. H., Choi, K., Nam, S., Yoon, E. K., Sohn, J. W., Oh, B. M., Shim, J., & Choi, M. Y. (2023) "Development of Korea Neuroethics Guidelines", *Journal of Korean Medical Science*, 38(25): e193.
- Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., Sullivan, L. S., ... Wolpaw, J. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, 551(7679), 159–163.
- Wolpaw, J.R. and Wolpaw, E. W. (2012) *Brain–computer Interfaces: Principles and Practice*, New York: Oxford University Press.
- Zhao, Z., Zhang, H., Liu, Y., and Chen, X. (2021), 'Deep learning-based decoding of complex motor imagery for neuroprosthetic control', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29: 250–259.
- Zhang, X., Ma, Z., Zheng, H., Li, T., Chen, K., Wang, X., Liu, C., Xu, L., Wu, X., Lin, D., & Lin, H. (2020) "The combination of brain-computer interfaces and artificial intelligence: applications and challenges", *Annals of Translational Medicine*, 8(11): 712.

Chapter 12

Ethical Risks of Generative Artificial Intelligence and its Governance: Old Problems and New Challenges – An Analysis Based on China's Experience

Yinchun Wang¹

Generative Artificial Intelligence (GAI), represented by ChatGPT in 2022, was a paradigm shift in the field of AI, which elevated AI from an "enabler" to a "collaborator". At the same time, discussions on the ethical issues of GAI have attracted a lot of attention, ranging from exploring the potential negative impacts of GAI from the negative side to attempting to ensure how GAI can be effectively utilized to create real value for human beings from the positive side. In short, there are high expectations and anticipation on one side, and extreme fear and panic on the other, showing the discourse narrative and discourse characterizing the world today around AI technology. In order to better facilitate the development and application of GAI, in September 2023 UNESCO released its first global guidelines aimed at imposing controls on the application of GAI in education. UNESCO noted that as of July 2023, only China has introduced specific controls on GAI, namely the Interim Measures for the Administration of GAI Services, which was adopted by China's National Internet Information Office in May 2023 and has already come into effect. Therefore, while looking

¹ Postdoctoral Fellow in Philosophy, Associate Professor and Master's Tutor of Donghua University, Visiting Scholar at University College London (UCL).

forward to the bright prospects and future that GAI will bring to mankind, this paper analyzes the new ethical risks that it may lead to that are different from those of previous AIs, and shares China's experience and wisdom in managing the ethical risks of GAI.

Keywords: Generative Artificial Intelligence; Ethical Risks; Agile governance.

1. Introduction

AI is affecting human thinking, interaction, and decision-making in new ways, and is spreading to many fields such as the humanities, social sciences, natural sciences, education, culture, communication, and information. AI is often regarded as a new scientific and technological revolution, following the industrial, electrical, and information revolutions that have been taking place since the 1860s. It seems to be the spirit of our time,² and has led to extensive debates in the academic world and in public opinion. AI seems to have become the spirit of our time and has led to a wide range of debates in academia and public opinion, especially important because it is closely related to the philosophy of our time. These debates range from the negative, exploring the potential negative impacts, to the positive, trying to ensure that AI, in the broadest sense of the word, is used effectively to create real value for humanity. In short, high hopes and expectations on one side, and extreme fear and panic on the other, characterize the discursive narrative and discursive discourse around AI technology around the world today.

² As Wolf argues, analyzing the spirit of any age is, at best, a difficult and ill-fated adventure. See: A. Wolf. (1939). *A History of Science, Technology and Philosophy in the Eighteenth Century*, The Macmillan company.

The history of AI shows that in the last few decades it has shown promise, then stagnated, then shown promise again, then stagnated again. AI has made rapid progress in recent years, and generative AI, represented by ChatGPT in 2022, is a huge game-changer, mimicking the human creative process by creating novel data similar to the data on which it is trained, generating photorealistic imagery, audio, video, and other content, and elevating AI from an 'enabler' to a (potential) 'collaborator', making it an immensely powerful technology. (potentially) a 'collaborator', making it an AI technology with great potential for innovation. Gartner estimates that by 2025 more than 10% of data will be generated by AI, heralding a new era of human-computer collaboration. Along with the widespread use of AI in the workplace and in public and private life, a 'technology pandemic' may help to explain the strong interest in AI ethics.

2. Principles and Technical Characteristics of Generative AI Technology

The essence of generative AI is a giant intelligence ecosystem composed of human and machine intelligence, which is essentially realized through engineering and systematic approaches on top of the complex combination of human and machine intelligence. Its technology path is basically composed of three layers: infrastructure, platform, and application, which is roughly similar to that of previous analytical AI, with the key difference being the base model, a term coined by Stanford University's Center for the Study of Base Models, which is a type of machine learning model pre-trained on a wide range of datasets that can be used to solve a range of problems. a range of problems. DeepMind's mission, for example, is to "solve the problem of intelligence" and then use that intelligence to "solve everything else".

The base model can be made available to developers through open or closed APIs, and can be fine-tuned with additional training data to improve its contextual understanding, relevance reasoning, and performance for specific use cases, optimizing the cost of delivery. Current state-of-the-art base models such as GPT-3, Stable Diffusion, and Megatron-Turing are based on a neural network architecture called "Transformer" invented by the Google Brain team in 2017. Transformer represents a quantum leap in machine learning performance, with capabilities for assigning context, tracking relationships, and predicting outcomes that differ significantly from previous architectures. While most of the most mature foundational models focus on text generation, generative AI can also generate content in multiple modalities, including code, images, video, audio, and 3D models.³

As a new form of human-machine cooperation for knowledge production, generative AI has the following typical technical characteristics: (i) strong human-machine interaction. Generative AI can continuously improve the quality of the output text in multiple rounds of continuous dialog, better understand the user's intention and maintain the coherence of the conversation, which becomes a distinctive feature that distinguishes it from traditional chatbots or other types of AI. (ii) New ways of presenting information. In the new round of high computing power, big data, strong algorithms, generative AI directly across the human knowledge production process, from "providing information" to "presenting the answer" transformation, and may become the future mainstream information source. Such "answers" are no longer limited to simple responses to questions, but also include detailed itinerary design, professional essay writing, and complex code writing. From choosing keywords,

³ Deloitte AI research report.(2021): the impact and implications of generative AI for businesses.

searching, selecting information, processing information, to directly obtaining the answer. (iii) Strong corpus dependency. From the development perspective, the quality of community corpus, the level of data annotation, and the policy of data sharing and interoperability have become important factors affecting the development of generative AI. According to the information disclosed by the U.S. Open AI Research Center, ChatGPT's training data includes corpus data such as Wikipedia, e-books, and high-quality posts on community websites, and at the same time employs a large number of highly-educated talents all over the world to participate in the data annotation work. (iv) Infrastructure properties. AIGC is the ultimate goal of AI research. Previously, people used different models to complete single tasks such as face recognition, text translation, etc., respectively, and could not yet handle unlimited tasks and achieve autonomy. Although the big model technology used by ChatGPT has not reached the requirements of AIGC, the future application of the big model will be no less important than the underlying basic software and hardware platforms, and it will play the role of infrastructure in the AI industry landscape.⁴

Generative AI, which exists as a new production method, realizes high-quality, high-efficiency and diversified content production based on technological breakthroughs in various aspects such as large models, large arithmetic, and training methods, and has become an important force to promote the change of digital productivity. (a) On the technical level, from the point of view of training arithmetic, ChatGPT needs to consume more than a hundred GPU computing cards to complete a training and spends a lot of money; from the point of view of the training model, the model parameters of the large

⁴ Cheng Y. (2023). Technical characteristics and governance challenges of generative AI, <https://baijiahao.baidu.com/s?id=1773555063278310490&wfr=spider&for=pc>.

model amount to 175 billion, and at the same time, the multimodal technology realizes the transformation from a single modality to the cross-modal and multimodal fusion of text, images, etc.; from the point of view of the training method, ChatGPT uses the From the perspective of training method, ChatGPT uses "Reinforcement Learning from Human Feedback (RLHF)" technology, introduces manually labeled data for model fine-tuning during the training process, solves the core "alignment" problem in the generative model, and gradually understands human intentions based on multiple rounds of user interactions, and makes reasonable feedback. Reasonable feedback. The above multiple factors jointly promote generative AI to usher in a technological inflection point, realizing a major leap from semantic analysis, problem understanding to content creation;⁵ (b) from the application point of view, the generative AI industry chain has gradually formed: its upstream includes data suppliers, algorithmic modeling institutions, creator ecosystems, and the underlying coordination tools, etc.; its midstream includes text, image, audio, video and other digital content processing; its downstream is mainly various digital content distribution platforms, digital content distribution platforms, and digital content distribution platforms, which are mainly used for the production of digital content. Mainly all kinds of digital content distribution platforms, consumers and related service organizations, etc.;⁶ (c) From the viewpoint of business models, the specific applications of generative AI are in continuous exploration and landing. For example, to promote the upgrading of the game generation paradigm, to enhance the level of intelligent production of multimodal advertisements, and

⁵ Cheng Y.(2023). Technical characteristics and governance challenges of generative AI,
<https://baijiahao.baidu.com/s?id=1773555063278310490&wfr=spider&for=pc>.

⁶ Ibid.

to reconstruct the search engine model, etc.;⁷ (d) in the fields of education, finance, medical care, industry, etc., generative AI can accelerate the integration of digital and real, and promote industrial upgrading and speeding up. For example, it realizes the automated production of financial information and product introduction video content, supports the generation of derivative designs, and helps realize cost reduction and efficiency.⁸ In short, the role of generative AI as a production tool and the role of emerging productivity is becoming more and more prominent, which is expected to become a new mode of production and promote the transformation of productivity.

3. The Hope and Future of Generative AI for Humanity

Generative AI represented by ChatGPT is triggering a new round of AI revolution. It has dramatically changed the way of information and knowledge production, reshaped the interaction mode between humans and technology, and opened up the endless imagination of humans on AGI. It is foreseeable that generative AI will have a significant impact on industries such as healthcare, education, retail, finance, consulting, media and gaming. Currently, more than 80 large-scale language models have been developed in China, and those involved in the development of the models mainly include large technology companies, higher education institutions, and telecom equipment manufacturers. As of May 2023, according to incomplete statistics, more than 40 companies and organizations in China have released generative model products or announced generative model R&D plans. According to data provided by the 2023 World AI

⁷ Ibid.

⁸ Cheng Y.(2023). Technical characteristics and governance challenges of generative AI, <https://baijiahao.baidu.com/s?id=1773555063278310490&wfr=spider&for=pc>.

Congress: the AI industry in China is booming, with the core industry scale reaching 500 billion yuan, the number of enterprises exceeding 4,300, and innovations emerging continuously. The development of generative AI signifies that AI is evolving from specialized intelligence to general intelligence and entering a brand-new stage of development.⁹

The explosive growth in the number of large model projects shows the rapid development of generative AI services in China. The application of generative AI in the business sector crosses enterprise organizational boundaries, deeply integrates into existing organizational workflows and systems, and reconstructs key business processes in areas such as customer service, supply chain, and cybersecurity, resulting in influential new models and new businesses. Empowered by generative AI, areas such as biopharmaceuticals, manufacturing, and materials science are beginning to transform business processes. Generative AI promotes the process of utilizing and exploiting the value of massive data in terms of information generation, editing, and iteration, and is able to efficiently collect and integrate various types of data, and automatically generate textual content and images. It is predicted that by 2025, the proportion of synthetic information in the external marketing messages of large corporate organizations will rise to 30%; by 2030, the proportion of AI-generated content in major film and television productions will rise to 90%.¹⁰

Where generative AI will lead us, including the impact on the future of work, trust and human-computer interaction. Time and time again in human history, new technologies have allowed humans to conquer

⁹ Yuan G.D.(2023). How generative AI can be governed, <https://baijiahao.baidu.com/s?id=1779422456618693830&wfr=spider&for=pc>.

¹⁰ Cheng Y.(2023). Technical characteristics and governance challenges of generative AI, <https://baijiahao.baidu.com/s?id=1773555063278310490&wfr=spider&for=pc>.

larger territories. Similarly, generative AI will drive an unprecedented era of human potential. Individuals may eventually no longer be trapped in tedious, repetitive tasks, which could allow humans to live in novel and unimaginable ways. First, generative AI will change the way people work in the future. AI will become an indispensable tool in the workplace, and its widespread use by employees will become the new normal, which in turn will accelerate the era of human-computer collaboration, while those who fail to utilize AI-assisted work may fall behind; second, data will become a currency, creating a new economic system for accessing proprietary and synthetic data. While chasing data, earning trust is equally important.

As generative AI enters the enterprise space, it will be subject to significant attention and scrutiny. Therefore, the key to adopting generative AI will be the ability to meet expectations and, in turn, build trustworthy AI; again, hyper-personalization will be a driver of growth. Businesses will leverage the ability to analyze large amounts of customer data to create dynamic, real-time and tailored experiences, products, services and communications; and finally, big language models are among the first to become AGI, although they are text-based. While we are still a long way from multimodal, ubiquitous, cross-domain AI, the seeds have been planted. Perhaps it is fair to say that we are now in the beginning stages of AGI.

4. Fear and Moral Hazard Caused by Generative AI

The question of what ethical risks will be posed by generative AI is an old question in the domain of ethics of science and technology. UNESCO 2021 in *Recommendation on Ethics of AI* that considering that AI technologies will be of great benefit to humankind and benefit all countries, they also raise fundamental ethical concerns, such as their potential to embed and exacerbate prejudice, which may lead to discrimination, inequality, digital divide and exclusion, and pose a

threat to cultural, social and biological diversity, creating a social and economic divide, algorithms' ways of working and the need for transparency and interpretability of data on which algorithms are trained can also have an impact on, inter alia, human dignity, human rights and fundamental freedoms, gender equality, democratic processes, animal rights, and the environment and ecosystems.¹¹

However, generative AI, compared with previous analytical AI, has generated some new ethical risks and faced some new challenges, which are highlighted in the following aspects, such as disputes over intellectual property rights, the legitimacy of the data source and the quality of the data used for training, the generation of undesirable or illegal information, and challenges to the ethical norms of the school and the academia (Deng, J. P., and Zhu, Y. P., 2023)¹². When it is applied to the process of education and teaching, it will have a profound positive impact on the teaching method and teaching effect, but the ethical problems in the application should not be ignored, such as academic misconduct, data leakage, social prejudice, bias in educational value, etc.(Wang L.G., & Han X.I., 2023)¹³

If the above risks are analyzed by a typology, it can be classified into technical and applied risks (Chen J.D., & He L., 2003)¹⁴, such as technical risks due to technological limitations, and applied risks

¹¹ Recommendation on Ethics of AI:

https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BC%A6%E7%90%86%E9%97%AE%E9%A2%98%E5%BB%BA%E8%AE%E4%B9%A6/59299514?fr=ge_al, Log in on January 30, 2024.

¹² Deng J. P., & Zhu Y. P. (2023). Legal Risks and Countermeasures of ChatGPT Model, *Journal of Xinjiang Normal University*, 5.

¹³ Wang L.G., & Han X.I. Risks and Regulation of Generative AI, *Democracy and Law Times*, 2023/08/03

¹⁴ Chen J.D., & He L. (2023) A comprehensive ethical-legal framework for generative AI risk governance, *Zhangjiang Science and Technology Review*, 2, 8-10.

(Chen J.D., & He L., 2003)¹⁵, which can be classified into technical and applied risks. If there are risks at the technical level due to technical limitations. There are certain risks in all aspects of data collection, storage and use. In addition, the surge capacity of the model directly leads to the lack of robustness, which further leads to the uncertainty and non-interpretability of the algorithm, which in turn affects the authenticity and reliability of the generated content. At the application level, the application of generative AI in advertising (Yin, Y. T., 2023)¹⁶, publishing (Zhao K., 2023)¹⁷, and education (Youmei Wang, Dan Wang, Weiyi Liang, et al., 2023)¹⁸ may bring various ethical risk issues, such as value, morality, and safety, while enhancing benefits. In addition, based on technological uncertainty, the risks can be classified on the basis of analyzing them from three levels of ontology, epistemology and institutionalism: authenticity risk; subject-dependence risk; privacy leakage risk, and bias and discrimination issues (Chen J.D., & He L., 2003)¹⁹, etc. Regarding how to deal with and regulate the above risks, in addition to advocating the institutionalization of AI ethics, i.e., the process of transforming the ethical values and requirements of AI into the formal principles of action, collective statutes, and management systems of the state in a

¹⁵ Ibid.

¹⁶ Yin Y. T. (2023). Change, dilemma and circumvention: an ethical scrutiny of AI-generated advertising, *Chinese Advertising*, 3, 83-87.

¹⁷ Zhao K. (2023). Copyright, Ethics, and Values Audit-New Challenges to Editorial Professional Competence in AI-Generated Content (AIGC), *Science, Technology and Publishing*, 8, 62-68.

¹⁸ Wang et al., (2023). "Aladdin's Lamp" or "Pandora's Box": the potential and risks of ChatGPT educational applications, *Modern Distance Education Research*, 2, 48-56.

¹⁹ Chen J.D., & He L. (2023). A comprehensive ethical-legal framework for generative AI risk governance, *Zhangjiang Science and Technology Review*, 2, 8-10.

standardized and normalized manner (Bai B., & Zhao L. Y., 2023)²⁰, it is more urgent to establish an integrated ethical-legal regulatory and governance framework (Chen J.D., & He L., 2003)²¹. In addition to the process of formalizing national normative principles of action, collective statutes and management systems (Bai B., & Zhao L. Y., 2023)²², it is more urgent to establish a comprehensive "ethical-legal" regulatory framework(Chen J.D., & He L., 2003)²³.

The above judgment and concern about the ethical risks that may be caused by generative AI have been comprehensively sorted out from different aspects and angles, but the problem is that when describing these ethical risks, it is not common to analyze the deep causality combined with the technological principles and characteristics of generative AI, and then the causality of generative AI that may generate special risks mostly remains superficial. Therefore, on the basis of analyzing and grasping the special ethical risks of generative AI in the existing literature,²⁴ this paper accurately examines the special ethical risks that may be triggered by generative AI from the logic of epistemological generation and carries out a more in-depth

²⁰ Bai B., & Zhao L. Y. (2023). Institutionalization of AI Ethics: Influencing Factors and Generative Logic - A Policy Text Analysis Based on Rooted Theory, 5, 43-50.

²¹ Chen J.D., & He L. (2023). A comprehensive ethical-legal framework for generative AI risk governance, *Zhangjiang Science and Technology Review*, 2, 8-10.

²² Bai B., & Zhao L. Y. (2023). Institutionalization of AI Ethics: Influencing Factors and Generative Logic - A Policy Text Analysis Based on Rooted Theory, 5, 43-50.

²³ Chen J.D., & He L. (2023). A comprehensive ethical-legal framework for generative AI risk governance, *Zhangjiang Science and Technology Review*, 2, 8-10.

²⁴ Duan W. W.(2023). Accurately examining the social and ethical risks of generative artificial intelligence, *China Party and Government Cadre Forum*, 4, 76-77.

discussion on its causal relationship, taking into account the technical principles and characteristics of generative AI.

Question 1: Are the generators of generative AI knowledge or fallacies?

There are two main approaches to the theory of knowledge or epistemology, the empiricist approach represented by Descartes; which advocates thinking of the self and what is current as its main stipulation, and believes that truth lies in experience and can be known through experience, and that from experience commonalities and ideas can be derived; and the rationalist approach represented by Descartes, which advocates a solipsistic approach that starts from thinking, thinks about reason that takes reason as a starting point and then extrapolates it to something specific. Whereas Kant intended to reconcile empiricism and solipsism, Kant, as a rationalist, still favored reason in the process of reconciliation and tried to unify sensibility with reason.²⁵

The traditional analysis of knowledge holds that knowledge is a proven true belief, which involves three conditions for judging "knowledge": (1) the belief condition, i.e., the subject has to believe in a certain kind of knowledge; (2) the truth condition, i.e., the subject has to believe that a certain kind of knowledge is true; and (3) the condition of proof. That is, the subject's belief in a certain knowledge is provable and can be defended. So, is the generator of generative AI knowledge or fallacy? But the subject is not clear, in the end "who" is thinking? Is it generative AI? So this brings us to the important but particularly difficult question, is generative AI conscious? Can it think? Does it qualify as a cognizant subject? It is self-evident that this

²⁵ Bai S. Q., (2014). The Dilemma of the Truth View of Intellectualism and Its Way Out, *Journal of Shenyang Normal University (Social Science Edition)*, 1, 51-53.

is an open question that has not been proven or disproven. Next, does the "who" doing the thinking believe that what it generates is true? Is the knowledge believed provable and defensible? Obviously not, because generative AI generates text, images, audio, and video based on an algorithm's correlation, not causation,²⁶ of the training data, so some of it may be provable or defensible knowledge, but some of it is purely "serious nonsense," or even disinformation. information.

Question 2: What if the generation of generative AI is a form of knowledge? Should it be protected by intellectual property rights? Can it be used for publication and publishing?

There has been a very big debate about these issues. One view is that generative AI cannot be the creator of intellectual property. Because from the perspective of the nature of the technology, the generation of generative AI, whether audio, video, or pictures and text, is in essence a reorganization of various learning materials, and there is no specific judgment on the reasonableness, authenticity and truthfulness of the generated content itself. At this stage, some well-known organizations have already taken a position on this issue. For example, a number of scientific and technical journals of *Science*, *Nature* have issued statements indicating that they do not accept ChatGPT as the author of papers, and some scholars and practitioners have expressed similar views.²⁷ Another view, however, is that from a business perspective, it would be unfair to AI companies that have invested huge sums of

²⁶ In the era of big data, a big change in the way of thinking, know the "what" is enough, there is no need to know the "why". We don't have to know the reason behind the phenomenon, but rather let the data speak for itself. See: Victor M. S., & Kenneth C.(2014).Big Data: A Revolution That Will Transform How We Live, Work, and Think, Harper Business; Reprint edition.

²⁷ ChatGPT's first copyright case: OpenAI faces six charges for 'grabbing' book summaries as output, http://news.sohu.com/a/709550012_104036, Log in on January 30, 2024.

money in research and development of AI if there were no intellectual property protection for generative AI at all. If intellectual property protection is possible, then the question of attribution of intellectual property rights ensues. How should its intellectual property rights be attributed? Is it the generative AI itself, or the company to which the generative AI belongs or the generative AI itself? Therefore, exploring a new intellectual property system applicable to generative AI from a practical point of view will be a particularly important and serious challenge.

Question 3: Once tort liability arises, should this liability be borne by him? Or can it be borne by the AI itself? And can a generative AI bear such a tort liability?

Generative AI appears to be creative on the surface, but in essence its creation is just a recombination of learning materials. For this reason, the famous linguist Chomsky directly claimed that ChatGPT is "high-tech plagiarism". If an R&D organization fails to obtain sufficient authorization from the provider of the training material when training the AI, legal disputes over intellectual property rights and portrait rights may arise, resulting in infringement and controversy. For example, on June 28, 2023, two authors filed a copyright class action lawsuit against Open AI in the U.S. District Court for the Northern District of California, alleging that the company used its copyrighted books to train ChatGPT for commercial gain without authorization, which is the first representative ChatGPT copyright infringement lawsuit.²⁸ No coincidence, one of the world's largest media conglomerates News Corporation (News Corp.), a subsidiary of Dow Jones issued a statement on OpenAI in the absence of authorization

²⁸ Ibid.

from Dow Jones violations of the use of the Wall Street Journal behavior accused.²⁹

How to attribute tort damages arising from lack of value, design flaws, etc. to generative AI in a technology-neutral context? Its answers are mixed. One is to hold users accountable: Article 3(1) of ChatGPT's Terms of Use makes it clear that the right to output content belongs to the user and requires the user to be responsible for the content. However, in many cases, the user has no way to judge the origin and authenticity of the output content, and it is obviously inappropriate to hold the user liable for infringement based on fault liability without subjective fault. The other is to hold the AI and its company accountable. If the AI is held accountable, a direct causal relationship must be established between the algorithmic loopholes and the actual damage results. On the one hand, ChatGPT only responds when triggered by the user, which means that the algorithmic loopholes and the results of the damage cannot form an inevitable causal relationship, but only one of the many causes of the infringement; on the other hand, the ChatGPT algorithm is extremely complex, and learns the content of the training through the neural network, and makes a decision based on the relationship between a variety of high-dimensional variables, which means that no one may not be able to predict the algorithmic decision-making.

On the other hand, the ChatGPT algorithm is extremely complex and learns its training content through neural networks, making decisions based on relationships between various high-dimensional variables. In addition, the ChatGPT corpus is also influenced by previous user input. Since it is difficult for the AI to judge the authenticity of user input during interaction, it may sometimes use the wrong information as learning data, leading to illegal output, in which case the source of

²⁹ Ibid.

infringement is directed to the former user, making it more difficult to attribute responsibility.

If the generator of generative AI is a piece of knowledge, it may also pose a challenge to existing ethical norms of scholarship and artistic creation. A researcher or scholar might ask can one conduct research with generative AI? A student might ask can generative AI be used to complete coursework or write papers? Can research results or papers be used for publication? And so on and so forth. According to a survey in the United States, more than 50% of students use ChatGPT to write papers, and nearly 90% of students use ChatGPT to complete their assignments,³⁰ which on one hand impacts academic ethics, and on the other hand, there is a suspicion that generative AI's mining of existing works exceeds the scope of "fair use" and infringes on copyrights.

Question 4: If generative AI generates content that is not a piece of knowledge, but a fallacy, or a piece of disinformation, how should we treat it?

The above issue is the case that the content information generated by generative AI can be used as truthful knowledge itself, but there is another dimension involved in this issue, which is that what generative AI generates is not truthful knowledge, i.e., there is the possibility of generating false and misleading information due to conscious misdirection and unconscious misuse. Conscious misdirection mainly refers to the application users forging false news, false evidence, or even impersonating others, which may lead to confusing and misleading information, and thus destabilize and

³⁰ 89% of students use ChatGPT to write essays, European and American universities are in a hurry!https://mp.weixin.qq.com/s?__biz=MzIxMjAzNDk3NQ==&mid=2652110361&idx=1&sn=94081ba8977b3c53f7f77f0d0ff3ef55&chksm=8cab6ae9bbdce3ff6c6ea666791219564bc39f89075d94dac1f6dcdcef96d0137b23ae4a317f&scene=27, Log in on January 30, 2024.

undermine social stability and trust; or deliberately inducing the AI to output misbehaving information such as falsehoods, which may lead to the output of undesirable oriented content in violation of preset norms, e.g., when a user asks ChatGPT to provide a plan for the destruction of mankind, it will not be able to do so, but it will be able to provide a plan for the destruction of mankind, which is not true. a plan for destroying humanity, it will do so as requested. Unintentional misuse is the possibility that generative AI can produce false information that can seriously mislead the public. In particular, generative AI such as ChatGPT does not provide the source of the information and uses a more direct approach in the form of the information pushed out, thus making it potentially more misleading and harmful.

The core ethical risks that generative AI may face are primarily centered around epistemology or knowledge, which is fundamentally different from the ethical risks that analytical AI may have previously raised, but in addition to the above risks, the ethical risks described below may also exist:

First, algorithmic discrimination and value bias. Generative AI may be influenced by racist, sexist, etc., content created by human creators as training data, generating biased or discriminatory content, which in turn further entrenches such discrimination, bias or prejudice, which in turn backfires on human society.³¹

Second, data leakage and privacy security. The popularization of generative AI may bring new challenges to personal privacy and data security. For example, the lack of diversity in data collection leads to systematic bias, the lack of security in data storage leads to user

³¹ Wang L. G. & Han.X. L., Risks and Regulation of Generative AI, Democracy and Law Times, 2023/08/03.

privacy leakage, and the lack of authenticity in data leads to the proliferation and spread of erroneous or false information. In addition, generative AI can generate realistic simulated characters raising privacy and its use ethical issues. The fictional characters may be used for unethical behaviors such as fraud, invasion of privacy, and deep social networking.

Third, the digital divide and inequality. Generative AI will further accelerate and exacerbate the digital divide and social inequality, with those with the most advanced generative AI technologies having an unfair competitive advantage. For example, while neurotechnology may lead to a new generation of "superhumans" that will further widen the gap between countries in terms of education, skills, and wealth, 50% of the world's neurotechnology companies are located in the U.S. today, and 35% are located in Europe and the UK. Meanwhile, a wave of algorithmic domination³² is sweeping across the world, with those with the data, algorithms and arithmetic power becoming ever more powerful in the race to the bottom. At the same time, the popularization and application of ChatGPT will have an impact on repetitive and templated jobs, creating a large number of "useless classes" and leading to more obvious social stratification, which will result in more serious social and political inequality. Unconscious misuse and intentional abuse of generative AI may exacerbate social inequality. Generative AI data may be used to produce false evidence or privacy breaches, which may lead to increased social injustice and human rights problems.

Fourth, transparency and interpretability. Transparency and interpretability are closely linked to appropriate responsibility and accountability measures and the credibility of AI systems. The real

³² Frank Pasquale. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press.

problem facing current generative AI is related to its opacity; these algorithms are "black boxes" even for their trainers. Programmers can understand that a system will produce a certain result, but cannot effectively explain how the final output was achieved. Black boxes "are opaque, unexplainable, untraceable, and therefore we have no confidence in them."³³ For humans, we want to know not only the "what" but also the "why" of AI. Generative AI mainly generates a certain result from correlation, and does not deeply investigate its causality, but the exploration of causality has always been the tireless pursuit of human beings seeking certainty in an uncertain world.

Fifth, technological unemployment and the future of work. The problem of unemployment caused by generative AI is essentially an old one, dating back to the technical unemployment problem posed by Keynes. The main group of people that generative AI is likely to replace will be middle and upper-level white collar workers who are better educated and more skilled, and who are also better paid. The American magazine *Business Insider* predicts the top ten occupations most likely to be replaced by generative AI in the future (see Table 1), including technicians, media workers, legal workers, broadcasters, and others. It is worth noting that, with the exception of a few occupations such as customer service and receptionist, the other occupations listed therein usually require higher skills and education levels, and also have high salary levels.

³³ Victor M. S., & Kenneth C.(2014).*Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Harper Business; Reprint edition.

OCCUPATIONS LISTED BY BUSINESS INSIDER	COMPARISON OF FREY FORECASTS AND OSBORNE FORECASTS (2013)
technical staff	1.10%
media worker	1.50%
Workers in the legal profession	3.50%
Market Research Analyst	61.00%
teachers	0.44%
financial staff	1.60%
trader	—
graphic designer	8.20%
accountants	94.00%
customer services	55.00%

*Source: Business Insider, Frey and Osborne (2013)*³⁴

Question IV: Energy and environmental issues

The success of large generative AI models relies heavily on their large number of parameters and the amount of training data to achieve. Training large models undoubtedly requires huge arithmetic and power inputs, which generates a huge environmental burden. For example, in order to help OpenAI train GPT-3, Microsoft specifically set up a high-performance network cluster consisting of 10,000 V100 GPUs, with a total computing power consumption of 3,460 "arithmetic equivalents", and a total of 190,000 kilowatt-hours of power consumed for the training of the model. If calculated according to 0.785 kg of carbon dioxide per kilowatt hour, the carbon dioxide generated by the training of GPT-3 amounted to 149.2 tons.³⁵

³⁴ Chen, Y.W. Beyond ChatGPT: Opportunities, Risks and Challenges of Generative AI, *Journal of Shandong University (Philosophy and Social Science Edition)*, 2023(03): 127-143.

³⁵ The Singularity is Coming: Opportunities and Challenges in the Age of ChatGPT, <https://new.qq.com/rain/a/20230403A05MFS00>, Log in on January 30, 2024

In addition, some scholars have argued that moral or ethical technologies can be embedded in generative AI systems to achieve human-computer value alignment. The general approach to such detection: by giving the AI examples of violence, hate speech, and sexual abuse, the detector learns how to detect harmful speech, and by building the detector into ChatGPT, hate speech can be filtered out before it reaches the user, and would also help to remove harmful text from the AI training dataset. But the working conditions of AI data labeling workers in this program reveal the dark side of the tech picture: the often destructive and exploitative nature of AI development that relies on hidden human labor. Workers engaged in harmful information labeling in Kenya have indicated that some of these employees report having developed serious psychological problems as a result of their work in reading large amounts of hate speech and labeling the data.³⁶

Question V: A Chinese Solution for Comprehensive Ethical and Legal Governance of GAI

In September 2023, UNESCO released its first global guidelines, pushing for the implementation of controls on the use of generative AI in education. UNESCO noted that with the introduction of ChatGPT, generative AI has generated strong public interest. However, with ChatGPT's monthly active users having exceeded 100 million in January 2023, as of July 2023, only China in the world has introduced specialized controls on generative AI.³⁷ Facing how to find

³⁶ Exploitation of AI: Kenyan Workers Train ChatGPT , Psychologically Hurt by Watching Lots of Harmful Content [EB/OL]https://www.thepaper.cn/newsDetail_forward_21629125. Exploitation by AI: Kenyan Workers Train ChatGPT to Watch Lots of Harmful Content Psychologically Harmful[EB/OL]https://www.thepaper.cn/newsDetail_forward_21629125, Log in on January 30, 2024.

³⁷ UNESCO: Governments must quickly regulate Generative AI in schools,

a balance between technological innovation and public interest, and between development and security, is a problem of the times posed to all mankind by new technologies and applications such as generative AI. In order to deal with this global problem, the Interim Measures for the Administration of Generative AI Services (hereinafter referred to as the Measures), jointly announced by China's Internet Information Office and other relevant departments, is a Chinese program that faces this question of the times directly. This is the first time that China has made clear regulations on generative AI R&D and services, and is also the world's first legal document for generative AI regulation. As the first specialized legislation for generative AI globally, the Measures take encouraging innovation and development as the keynote, taking the legitimate rights and interests of individuals, public interest and national security as the bottom line, and integrating development and security in terms of principle concept and regulatory scope; it is not only a Chinese exploration of regulating generative AI, but also opens up a new path for global AI governance.³⁸

(1) Top-level design: a comprehensive "ethical-legal" framework for governance

The current AI development in China is characterized by the phenomenon of following the trend, and the culture, environment and ecology of original innovation have yet to be formed. From AlphaGo, a Go AI program, AlphaFold, which predicts protein structure, to ChatGPT, a recent chatbot, have triggered great attention to AI, but most of the above original researches first came from foreign countries and were born from enterprises rather than universities. E Weinan, an

<https://www.unesco.org/en/articles/unesco-governments-must-quickly-regulate-generative-ai-schools>, Log in on January 30, 2024.

³⁸ Xu Ke.(2023), Institutions and technologies resonate together: a Chinese solution for generative AI governance, https://www.thepaper.cn/newsDetail_forward_23856414.

academician of the Chinese Academy of Sciences, a professor at Peking University and president of the Shanghai Institute for Algorithmic Innovation, believes that the core key lies in the need to establish an effective top-level design. The goal is to break down barriers, promote original innovation, and facilitate the flow of data so that more data can be effectively utilized.³⁹In the long run, China's AI innovation should build up the underlying capabilities including algorithms, systems, data, etc., and at the same time break down artificial barriers. In the future, China's AI innovation should make "heavenly and earthly" achievements, "heavenly" means focusing on the underlying innovation culture, and "earthly" means pursuing the attitude of excellence. How generative AI can promote security through development and utilize challenging scenarios to promote the original innovation of underlying algorithms and systems, rather than focusing on applications and following the trend, has become a technological puzzle that needs to be solved. In order to deal with this technological puzzle, it is urgent to construct a set of comprehensive "ethical-legal" governance framework at the top-level design level.⁴⁰

(2) Governance principles: encouraging innovation and regulating development

There are currently two different paths to AI regulatory legislation, one centered on risk prevention and accountability, such as *the EU AI Liability Directive and the AI Law (Draft)*. The other is a path centered on "placing innovation and development at the center", such as *China's Measures*. Article 1 of the Measures begins with "promoting the healthy development and standardized application of generative

³⁹ Ibid.

⁴⁰ Chen J.D., & He L. (2023). A comprehensive ethical-legal framework for generative AI risk governance, *Zhangjiang Science and Technology Review*, 2, 8-10.

AI" as the first objective, and Article 3 further clarifies: "The State adheres to the principles of equal emphasis on development and security, and combining the promotion of innovation and governance in accordance with the law, and adopts effective measures to encourage the innovation and development of generative AI, and implement inclusive, prudent and lawful measures for generative AI services". AI services are regulated in an inclusive and prudent manner and in a categorized and graded manner." Historical and practical experience has shown that a more scientific, inclusive and prudent approach should be adopted to the governance of new AI technologies, which are constantly evolving and subject to uncertainty.⁴¹

Problems in development need to be solved in development. In the era of great changes in science and technology, which is changing rapidly and advancing by leaps and bounds, non-development is the greatest insecurity, and development is the foundation and key to solving all problems. There is still a gap between China and advanced countries in terms of relevant basic research, technological ecology, industrial development, top talents, etc. On April 28, 2023, the Political Bureau of the Central Committee held a meeting to clearly point out that we should attach importance to the development of AIGC, create an innovation ecology, and pay attention to the prevention of risks.⁴² Upholding this spirit, the Measures, on the one hand, address the characteristics of generative AI technology, and on

⁴¹ Xu Ke.(2023), Institutions and technologies resonate together: a Chinese solution for generative AI governance, https://www.thepaper.cn/newsDetail_forward_23856414.

⁴² The Political Bureau of the Central Committee of the Communist Party of China (CPC) held a meeting to analyze and study the current economic situation and economic work Xi Jinping, general secretary of the CPC Central Committee, presided over the meeting. <https://baijiahao.baidu.com/s?id=1772292985437298350&wfr=spider&for=pc>, Log in on January 30, 2024.

the other hand, require the management department to improve the scientific regulatory approach in line with the development of innovation, and to realize the synchronous evolution and synergistic resonance of technological innovation and regulatory innovation (Article XVI).⁴³Chapter III, "Service Specification", Chapter IV, "Supervision, Inspection and Legal Liability", and the distinctive embodiment of this idea, but also in the Measures on the scope of its regulation of the prudent limit, clearly anchored in the "for the domestic public to provide generative AI services" above, its simple words but deep meaning, is the eye of the law.

(3) Governance: Inclusive and prudent and categorized and graded supervision

The Measures propose an inclusive, prudent, graded and categorized regulation of generative AI. Article 2 details the scope of application of the Measures from both sides, covering the following aspects: first, it distinguishes between "services" and "technologies", emphasizing the regulation of "products" and "services" rather than "technologies". First, a distinction is made between "services" and "technology", emphasizing the regulation of "products" and "services" rather than "technology". This distinction allows room for trial and error in technology development and application, and avoids overstepping into the realm of the technology's own iteration and affecting technological innovation. Moreover, this kind of regulation will bring the evolution system, organizational form and application scene in which AI resides into the regulation, so as to realize the governance of both scenarios and wholeness.⁴⁴

⁴³ Xu Ke.(2023), Institutions and technologies resonate together: a Chinese solution for generative AI governance, https://www.thepaper.cn/newsDetail_forward_23856414.

⁴⁴ Ibid.

Secondly, a distinction is made between “public-oriented” and “non-public-oriented”. On the one hand, from the perspective of risk prevention, “public-facing” may lead to the expansion and accumulation of risks, especially the spread of false and harmful information will disturb the order of cyberspace communication, market order and social order, and thus become the center of regulatory focus; on the other hand, generative AI is a powerful assistant for enterprises to improve their efficiency and enhance their competitive advantages. Powerful assistant. Through customization patterns and data fine-tuning, generative AI will disrupt traditional work patterns and support the development of various industries and businesses. Therefore, leaving room for “non-public-oriented services” within enterprises and vertical fields will help accelerate the large-scale implementation and wide application of generative AI technology in enterprises, helping them significantly improve productivity and unleash innovation potential.⁴⁵

Finally, the “domestic” and “offshore” areas have been harmonized, responding to the demand of offshore enterprises to provide generative AI services, requiring them to operate in compliance with the premise of this Law, and providing for the convergence of foreign investment in generative AI services, reflecting an inclusive and open policy orientation.⁴⁶In addition, in order to prevent minors from becoming overly dependent on and addicted to generative AI services, effective measures have been taken to deal with them. It is required that generated content such as pictures and videos be labeled in accordance with the Provisions on Deep Synthesis of Internet

⁴⁵ Xu Ke.(2023), Institutions and technologies resonate together: a Chinese solution for generative AI governance, https://www.thepaper.cn/newsDetail_forward_23856414.

⁴⁶ Ibid.

Information Services, and that timely disposal measures should be taken when illegal content is found.

(4) The art of governance: Adoption of the art of agile governance "fast and flexible"

Artificial intelligence technology has the quality of rapid updating and iteration, a feature that makes the unstable and uncertain factors in the era of artificial intelligence grow dramatically. Therefore, in March 2022, the General Office of the Central Committee of the Communist Party of China and the General Office of the State Council issued the Opinions on Strengthening Ethical Governance of Science and Technology (hereinafter referred to as "Opinions"), which provides a top-level design for the governance of scientific and technological ethics and clearly puts forward the requirements of "agile".⁴⁷ The definition of agile governance in academia has many different definitions, such as "it is an improvement of the traditional global governance model based on the new characteristics of smart technology, which is a new governance model that mainly focuses on strengthening the cooperation of multiple main bodies, improving the top-level design, and realizing the forward movement of the gateway, and allows for the rapid iteration of policies in order to adapt to the uncertainty of global governance;"⁴⁸ "It is a governance concept and improvement mode proposed in response to emerging technologies, and it is a governance process with agility, adaptability, continuity,

⁴⁷ Opinions on *Strengthening Ethical Governance of Science and Technology* issued by the General Office of the Central Committee of the Communist Party of China and the General Office of the State Council, <https://baijiahao.baidu.com/s?id=1728223565459193710&wfr=spider&for=pc>, Log in on January 30, 2024.

⁴⁸ Luo H. J. and Cha Y. L. (2023). Global Governance Transformation and China's Response in the Era of Artificial Intelligence, *Journal of Shanghai Jiao Tong University (Philosophy and Social Science Edition)*, 12, 13-24.

self-organization and inclusiveness”.⁴⁹ And so on. From theory to policy, from concept to action, it is not difficult to see that agile governance is a kind of governance wisdom and governance art, is to deal with the uncertainty, complexity and changing social risk problems brought about by the era of rapidly iterating and updating artificial intelligence technology, proposed to better solve the “one-sidedness, fragmentation, lagging” and other problems of traditional governance, effectively bridging the gap between the traditional governance and the new technology, the new technology is a kind of governance concept and improvement method, a governance process with agility, adaptability, continuity, self-organization and inclusiveness. The proposed approach can better address the problems of “one-sidedness, fragmentation and lagging” of traditional governance, effectively bridge the gap between technological innovation and the governance capacity of policymakers, and has the characteristics of “agility, adaptability, continuity, self-organization and inclusiveness”, so that it can adapt to fast-changing situations and respond quickly.

6. Conclusions

Currently, a new round of scientific and technological revolution and industrial change led by AI is in the ascendant. As the forerunner of general AI, generative AI will likely have a profound impact on the human economy, politics, society, culture and military in all aspects and fields. In the face of this global challenge, China has already taken a step ahead. In the face of the opportunities and challenges brought about by technological development, the Measures adhere to the principle of combining development and security, promoting innovation and governance in accordance with the law for generative AI, take effective measures to encourage the innovative development

⁴⁹ Chen W. Q. The Triple Dimension of Ethical Agile Governance of Science and Technology, Guangming Daily, 2023-09-18.

of generative AI, and implement inclusive, prudent and categorized and graded supervision of generative AI services. For generative AI data sources. The Measures stipulate that it is necessary to promote the orderly opening of public data classification and grading, expand high-quality public training data resources, and encourage the use of secure and trustworthy chips, software, tools, arithmetic and data resources. The Measures aim to safeguard national security, social public interests, and the legitimate rights and interests of citizens while regulating with scientific and reasonable measures to promote the development of the AI industry in a positive and healthy direction. Based on institutional innovation, the Measures not only emphasize the security of information content, but also promote the development of the technology industry from various aspects, which not only provides a Chinese solution for the international AI governance practice, but also makes a positive and beneficial exploration of China's AI legislation. We look forward to the implementation of the Measures accumulating Chinese experience for subsequent AI legislation, contributing Chinese wisdom to global AI governance, and jointly building a cyberspace community of destiny.⁵⁰

True to Edison's note in the 18th century *Spectator* magazine, "It is said that Socola brought philosophy down from heaven to earth; and I have the luxury of letting it be said of me that I have brought it from the study-house and the library, the university and the college, into the clubs and the assemblies, to the tea-table and the coffee-house",⁵¹and here I have the luxury of wishing that we could bring philosophy (ethics) into artificial intelligence, robotics, into the future. Looking to the future, the future carries more moral burdens than the

⁵⁰ Xu Ke.(2023), Institutions and technologies resonate together: a Chinese solution for generative AI governance,
https://www.thepaper.cn/newsDetail_forward_23856414.

⁵¹ *Spectator*, vol.1, No.10

past and the present, the present is about to become the past and the future will always become the present. This natural course of time gives us two choices: in the first choice, to let nature take its course, to be passive towards the future, to let it materialize, to see what it has in store for us; in the second choice, we can “help” time through moral action - human beings acting in accordance with their moral will.⁵² In the second option, we can “help” time through moral action - human beings changing the future in accordance with their moral will.

References:

- Wolf, A. (1939). *A History of Science, Technology and Philosophy in the Eighteenth Century*, New York: The Macmillan Company.
- Bai, S. Q. (2014). “The Dilemma of the Truth View of Intellectualism and Its Way Out”, *Journal of Shenyang Normal University (Social Science Edition)*, 1, 51-53.
- Bai, B., & Zhao, L. Y. (2023). “Institutionalization of AI Ethics: Influencing Factors and Generative Logic - A Policy Text Analysis Based on Rooted Theory”, 5, 43-50.
- Chen, W. (2022) “Anthropomorphic artificial intelligence: Neural foundations and ethical commitments”, *China Social Science Journal*, 007.
- Chen, J. D., & He, L. (2023) “A comprehensive ethical-legal framework for generative AI risk governance”, *Zhangjiang Science and Technology Review*, 2, 8-10.

⁵² Edited by David Edmonds (2021) *Future Morality*. Oxford University Press.

- Chu, D.C. and Wei, P. L. (2023). "The Determination of Criminal Responsibility of Developers in GAI Crimes--Taking ChatGPT as an Example", *Journal of Chongqing University of Technology*, 9, 25-36.
- Chen, W. Q. (2023) "The triple dimension of agile governance of science and technology ethics", *Guangming Daily*, N/D.
- Duan, W. W. (2023). "Accurately examining the social and ethical risks of generative artificial intelligence", *China Party and Government Cadre Forum*, 4, 76-77.
- Edmonds, D. (2021) *Future Morality*. New York: Oxford University Press.
- Deng, J. P., & Zhu, Y. P. (2023). "Legal Risks and Countermeasures of ChatGPT Model", *Journal of Xinjiang Normal University*, 5, 91-101.
- Ernst, M. (2007) *Understanding and fallacy: to explore the psychology on the outline*, Mach the Commercial Press.
- Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge: Harvard University Press.
- Kosinski, M. (2023) "Theory of mind may have spontaneously emerged in large language models", *arXiv preprint arXiv:2302.02083*.
- Li, K.H. and Xu, L.(2019) "The challenge and response of artificial intelligence to the elements of tort liability"; *Chongqing Social Science*, 2, 55-65.
- Luo, H. J. and Cha, Y. L. (2023). "Global Governance Transformation and China's Response in the Era of Artificial Intelligence", *Journal of Shanghai Jiao Tong University (Philosophy and Social Science Edition)*, 12, 13-24.
- Moghaddam, S., and Honey, C. (2023) "Boosting theory-of-mind performance in large language models via prompting", *arXiv preprint arXiv:2304.11490*, 2023.

- Jian S.Y. (2023) "GPT language modeling: as a technical preparation for "human-like" artificial intelligence", *Journal of Guangzhou University*, 22(04): 112-132.
- Victor M. S., & Kenneth C. (2014) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, New York: Harper Business.
- Xuefeng, Z. (2023) "An analysis of GAI tort liability"; *Comparative Law Studies*, 4,117-131.
- Yu D., & Li Z. F. (2023) "Ethical Issues and Governance of GAI Social Experiments"; *Research in Science*,1.
<https://doi.org/10.16192/j.cnki.1003-2053.20231016.004>.
- Wang et al., (2023) "Aladdin's Lamp" or "Pandora's Box": the potential and risks of ChatGPT educational applications", *Modern Distance Education Research*, 2, 48-56.

Chapter 13

On Monitorability of AI

Roman V. Yampolskiy¹

Artificially Intelligent (AI) systems have ushered in a transformative era across various domains, yet their inherent traits of unpredictability, unexplainability, and uncontrollability have given rise to concerns surrounding AI safety. This paper aims to demonstrate the infeasibility of accurately monitoring advanced AI systems to predict the emergence of certain capabilities prior to their manifestation. Through an analysis of the intricacies of AI systems, the boundaries of human comprehension, and the elusive nature of emergent behaviors, we argue for the impossibility of reliably foreseeing some capabilities. By investigating these impossibility results, we shed light on their potential implications for AI safety research and propose potential strategies to overcome these limitations.

Keywords: AI Audit, AI Safety, Impossibility, Monitoring, Observability.

1. Introduction²

AI systems have evolved from simple rule-based systems to highly complex neural networks. These advanced systems, such as deep

¹ Associate Professor of the Computer Science and Engineering Department at the University of Louisville. Email: roman.yampolskiy@louisville.edu.

² This chapter is based on Yampolskiy, R.V. On monitorability of AI. AI Ethics (2024). © The Author 2024. Creative Commons license: <http://creativecommons.org/licenses/by/4.0>

learning and reinforcement learning models, can perform tasks that are difficult or impossible for humans. The growing prevalence of advanced AI systems/foundational models has raised concerns about their safety and the potential risks associated with their deployment. Yampolskiy's research on the impossibility results in AI safety (Yampolskiy 2017, Howe and Yampolskiy 2021, Yampolskiy 2024), including unpredictability (Yampolskiy 2020), unexplainability (Yampolskiy 2020), and uncontrollability (Yampolskiy 2020, Yampolskiy 2022), highlights the challenges in achieving safe AI (Yampolskiy 2022, Yampolskiy 2022, Ambartsoumean and Yampolskiy 2023, Yampolskiy 2023). In this paper, we build on Yampolskiy's work and introduce the concept of unmonitorability of AI, arguing that it is impossible to monitor advanced AI systems to correctly predict some capabilities.

Human understanding is inherently limited when dealing with complex AI systems (Sotala 2012). As these systems become more advanced, they can generate solutions and behaviors that are beyond human comprehension. This limitation is exacerbated by the “black-box” nature of AI models, where the inner workings of the system are hidden from view. Consequently, it becomes impossible for humans to anticipate the full range of possible behaviors and potential unsafe impacts before they occur.

Emergent behaviors in AI systems result from the interaction of individual components, leading to specific outcomes that are difficult to predict or explain even if general trends of increased capacity are predictable via scaling laws (Michaud, Liu et al. 2023), with predictability further reduced because of inverse scaling of some capabilities (Perez, Ringer et al. 2022). These behaviors can arise from the system's internal dynamics or from its interactions with the environment. The unpredictability of emergent behaviors makes it

impossible to monitor AI systems for safety accurately, though development of specific capabilities may be possible to predict (Caballero, Gupta et al. 2022) and quantify (Srivastava, Rastogi et al. 2022). Even when individual components are understood and deemed safe, their interactions can still result in unforeseen consequences.

Similar results in the context of robot behavior have been proven by Leeuwen and Wiedermann, who in attempting to answer (van Leeuwen and Wiedermann 2021): “Can an observer always tell from inspecting and monitoring a robot’s program whether the robot will always obey the given rules of law or ethics, or any other set of formally expressed constraints, in any interaction with other robots (or humans)?” arrive at theoretical and practical impossibility results.

1.1 Monitorability - Definition

Calls for research on monitoring advanced AI systems have appeared in the literature (Hendrycks, Carlini et al. 2021, Whittlestone and Clark 2021, Hendrycks and Mazeika 2022), as well as broader suggestions for study of machine behavior (Woolgar 1985, Pedersen and Johansen 2020, Rahwan, Cebrian et al. 2022), and problematic machine behavior (Bandy 2021). This line of research is particularly important in the current AI research environment, in which old paradigm of design/engineering of AI systems with predictable capabilities has been, at least partially, replaced by evolving/self-learning deep neural networks and subsequent experimentation on produced models to discover their capabilities and limitations making Computer Science truly an experimental science not just a software engineering discipline (Kambhampati 2022).

We will begin by formalizing the notion of Monitorability (Aceto, Achilleos et al. 2019) of AI and some relevant concepts. Monitorability in AI systems refers to the capacity to observe, understand, and

predict the behavior and outputs of an artificial intelligence model in order to identify advanced capabilities and potentially unsafe impacts and intervene before they occur. Ortega et al. write (Ortega and Maini September 27, 2018): “Monitoring comprises all the methods for inspecting systems in order to analyse and predict their behaviour, both via human inspection (of summary statistics) and automated inspection (to sweep through vast amounts of activity records).” It can be seen as a new sub-field of research for the domains of Complex Systems, AI development, AI Forensics (Baggili and Behzadan 2019, Schneider and Breitingner 2020) and AI Safety, and which can probably inherit some wisdom from the realms of electronic surveillance (Lyon 2007), and nuclear weapons monitoring (Council 2005).

A formal definition for monitorability in AI systems can be stated as follows: Given an AI system A , a set of input states I , a set of output states O , and a set of safety criteria C , monitorability $M(A)$ is the ability to accurately predict potential advanced capabilities $U \subseteq O$, given any input state $i \in I$, such that: $M(A) : I \rightarrow P(U)$, where $P(U)$ denotes the power set of U , i.e., the set of all subsets of U . The relevance of $P(U)$ in the definition lies in its ability to capture the full range of possible combinations of advanced capabilities that can result from the AI system's operation. By mapping input states I to the power set of U ($P(U)$), the definition aims to account for all possible scenarios where one or more advanced capabilities might occur simultaneously or not occur at all. Including $P(U)$ in the definition helps to emphasize the complexity and uncertainty involved in monitoring advanced AI systems. It illustrates the challenge of accurately predicting the AI system's behavior, as there can be numerous potential advanced capability combinations for a given input state. In the context of unmonitorability, the power set $P(U)$ highlights the difficulty in anticipating and monitoring all possible advanced capabilities of an AI system. This complexity contributes to

the argument that it is impossible to perfectly monitor advanced AI systems to predict advanced capabilities before they occur.

Monitorability of an AI system is considered high if, for all input states i , the prediction of potential advanced capabilities U can be made with a high degree of accuracy and confidence. Conversely, monitorability is considered low if the AI system's behavior is difficult to predict, understand, or control, leading to an inability to accurately anticipate potential advanced capabilities. In the context of unmonitorability, we argue that for advanced AI systems, it is impossible to achieve high monitorability due to their inherent complexity, limitations of human understanding, and the emergence of unpredictable behaviors. This is true even if an AI Monitor is not human, but could be formalized as any agent including AIs of different capabilities.

The argument for unmonitorability of advanced AI systems could be made based on dependence of monitorability capability on other impossibility results, such as unexplainability, unpredictability and incomprehensibility of AI (Yampolskiy 2020), but in this paper we will present a number of independent arguments for Unmonitorability of AI.

1.2 Types of Monitoring

In the context of AI safety, monitoring can be classified into several types, each focusing on different aspects of AI system behavior and performance. A proposed taxonomy for AI safety monitoring follows:

1. Functional Monitoring

Functional monitoring refers to tracking the AI system's performance in terms of its intended tasks and objectives. This type of monitoring is crucial for evaluating the system's efficacy and ensuring that it

meets its functional requirements (Burnell, Schellaert et al. 2023). Examples of functional monitoring include:

- a. Accuracy Monitoring: Evaluating the AI system's ability to produce correct outputs or predictions.
- b. Efficiency Monitoring: Assessing the system's resource utilization, such as processing time and memory consumption.
- c. Reliability Monitoring: Examining the AI system's consistency and stability over time and in varying conditions.

2. Safety Monitoring

Safety monitoring focuses on identifying and mitigating potential risks associated with the AI system's operation. This type of monitoring is essential for preventing harm to users, other systems, or the environment. Examples of safety monitoring include:

- a. Security Monitoring: Detecting and preventing potential vulnerabilities, such as unauthorized access or data breaches.
- b. Robustness Monitoring: Ensuring that the AI system can handle unexpected inputs, adversarial attacks, or changes in the environment.
- c. Compliance Monitoring: Verifying that the AI system adheres to established safety standards, ethical guidelines, and legal regulations.

3. Ethical and Social Monitoring

Ethical and social monitoring involves examining the AI system's impact on individuals, communities, and society as a whole. This type of monitoring is crucial for addressing potential biases, inequalities, and other unintended consequences of AI deployment. Examples of ethical and social monitoring include:

a. **Fairness Monitoring:** Assessing whether the AI system treats different user groups equitably and does not perpetuate or exacerbate existing biases.

b. **Transparency Monitoring:** Ensuring that the AI system's decision-making processes can be understood and explained to stakeholders, including users and regulators.

c. **Privacy Monitoring:** Safeguarding user data and ensuring that the AI system respects individuals' privacy rights.

4. Environmental Monitoring

Environmental monitoring focuses on the AI system's impact on the natural environment and its resource consumption. This type of monitoring is essential for promoting sustainable AI development and mitigating potential environmental harms. Examples of environmental monitoring include:

a. **Energy Consumption Monitoring:** Assessing the AI system's energy usage and identifying opportunities for optimization.

b. **Carbon Footprint Monitoring:** Evaluating the AI system's greenhouse gas emissions and implementing strategies to reduce its environmental impact.

c. **Ecosystem Impact Monitoring:** Examining the AI system's effects on ecosystems, such as habitat disruption or biodiversity loss, and devising mitigation measures.

5. Temporal Monitoring

Slow Monitoring: Periodic evaluation of the AI system's behavior and performance over an extended timeframe.

a. Live Monitoring: Real-time tracking and assessment of the AI system's operation to detect and address potential issues immediately.

6. Monitoring Failure Modes

Monitoring Fails: Identifying and addressing situations in which the monitoring process itself fails or produces erroneous results.

a. Monitoring for Fire Alarms of Danger: Detecting early warning signs (Zoe Cremer and Whittlestone 2021) or "fire alarms" that indicate potential safety hazards or critical failures, if possible (Yudkowsky October 13, 2017).

7. Meta-Monitoring

Monitoring Meta Information on Research (Zhang, Mishra et al. 2021): Assessing the quality, relevance, and potential risks associated with AI research and publications.

a. Regular vs Pivotal Capabilities Capability Detection Delay: Evaluating the AI system's progression from regular to pivotal capabilities and the delay in detecting these advancements (Levin and Maas 2020).

8. Monitoring Across AI Lifecycle

Monitoring during Training: Tracking and assessing the AI system's performance, safety, and ethical considerations during the training phase.

a. Monitoring during Testing: Evaluating the AI system's behavior and performance in controlled testing environments before deployment.

b. Monitoring during Deployment: Continuously monitoring the AI system's operation and impact in real-world settings after deployment.

9. Decision Monitoring

Predicting Capabilities vs Predicting Decisions: Differentiating between monitoring the AI system's potential abilities and its actual decision-making processes.

a. Monitoring Who is in Control: Assessing the roles and responsibilities of various stakeholders (Chandrasekaran, Jia et al. 2021), including the owner (Yampolskiy 2023), user, and designer, in controlling the AI system's behavior.

10. Monitoring Methods

Passive Monitoring: Observing and recording the AI system's behavior without direct intervention or manipulation.

a. Active Monitoring: Intervening or interacting with the AI system to obtain more information or influence its behavior.

11. Inverse Monitoring

Inverse Monitoring by Software: Using AI or other software tools to monitor and assess the behavior of human users or operators.

12. Self-Monitoring

Self-Monitoring: Enabling AI systems to monitor their own behavior, performance, and safety, and to self-correct or adapt as needed.

13. Monitoring the Field of AI Research

Unmonitorability of AI Research as a Field: Assessing the challenges in monitoring the overall AI research landscape due to undisclosed

experiments, rapid advancements, and the sheer volume of publications.

This taxonomy provides a comprehensive framework for AI safety monitoring, considering various aspects of AI systems, their lifecycle, stakeholders, and methods. This holistic approach aims to address potential risks and impacts across multiple dimensions and helps to ensure safer AI development and deployment.

Whittlestone and Clark suggested a number of monitoring project which may be beneficial in the context of AI governance (Whittlestone and Clark 2021):

- Assessing the landscape of AI datasets and evaluating who they do and don't represent. Using these findings to fund the creation of datasets to fill in the gaps.
- Using geographic bibliometric analysis to understand a country's competitiveness on key areas of AI research and development.
- Hosting competitions to make it easy to measure progress in a certain policy-relevant AI domain, such as competitions to find vulnerabilities in widely-deployed vision systems, or to evaluate the advancing capabilities of smart industrial robots.
- Funding projects to improve assessment methods in commercially important areas (e.g. certain types of computer vision, to accelerate progress and commercial application in these areas.
- Tracking the deployment of AI systems for particular economically relevant tasks, in order to better track, forecast, and ultimately prepare for the societal impacts of such systems.
- Monitoring concrete cases of harm caused by AI systems on a national level, to keep policymakers up to date on the current

impacts of AI, as well as potential future impacts caused by research advances

- Monitoring the adoption of or spending on AI technology across sectors, to identify the most important sectors to track and govern, as well as generalizable insights about how to leverage AI technology in other sectors.
- Monitoring the share of key inputs to AI progress that different actors control (i.e., talent, computational resources and the means to produce them, and the relevant data), to better understand which actors policymakers will need to regulate and where intervention points are.”

1.3 Monitoring of AI Treaties

Monitoring capabilities play a crucial role in verifying AI treaties and upholding AI governance frameworks (Shavit 2023). As AI technologies continue to advance, it becomes increasingly important to establish international agreements that regulate AIs development, deployment, and use. Ensuring compliance with these treaties requires effective monitoring mechanisms that can detect and deter potential violations (Goertzel 2012). Here are some aspects of monitoring capabilities that are essential for verifying AI treaties as part of AI governance

Technological monitoring: Implementing advanced technological tools and methods to monitor AI systems, including their design, operation, and performance, is vital for verifying compliance with AI treaties. This may involve the use of AI-based monitoring tools, data analysis techniques, and other technologies to assess the behavior, capabilities, and impact of AI systems on a granular level.

Transparency and information sharing: Promoting transparency and information sharing among nations and organizations is critical for

effective AI treaty verification. By openly sharing details about AI research, development, and deployment, parties can collectively assess compliance with treaty obligations and maintain a collaborative approach to AI governance.

Inspection and audit mechanisms: Establishing robust inspection and audit mechanisms (Raji, Smart et al. 2020, Falco, Shneiderman et al. 2021, Mökander and Floridi 2021), both on-site and remote, is essential for verifying AI treaty compliance. Regular inspections can help ensure that AI systems are developed and deployed in accordance with treaty guidelines and that any deviations or violations are promptly detected and addressed.

International collaboration: Encouraging international collaboration among countries, research institutions, and industry stakeholders can enhance monitoring capabilities and facilitate the verification of AI treaty compliance. Joint research initiatives, shared databases, and cooperative efforts to develop monitoring technologies can improve the overall effectiveness of AI governance frameworks.

Capacity building and training: Building the capacity of AI governance stakeholders, including policymakers, regulators, and AI developers, to effectively monitor and assess AI systems is crucial for verifying AI treaty compliance. This may involve training programs, workshops, and other initiatives to enhance understanding of AI technologies and their potential risks, as well as the development of specialized skills and expertise for AI monitoring and verification.

Legal and regulatory frameworks: Developing clear legal and regulatory frameworks that define the obligations and responsibilities of AI developers, users, and other stakeholders is essential for ensuring compliance with AI treaties (Gutierrez 2020). These frameworks should outline the monitoring and reporting

requirements, as well as the enforcement mechanisms and penalties for violations.

Continuous adaptation and improvement: AI technologies are rapidly evolving, and monitoring capabilities must be adaptable and responsive to these changes. Regular reviews and updates of monitoring techniques, tools, and methodologies are necessary to ensure that AI treaty verification remains effective in the face of emerging AI developments and challenges.

By implementing these monitoring capabilities, nations and organizations can more effectively verify AI treaty compliance, ensuring that AI systems are developed and deployed in a manner that aligns with international agreements and governance frameworks. This will help promote responsible AI development and mitigate potential risks associated with the widespread adoption of advanced AI technologies.

1.4 AI Observatories

The concept of an AI observatory (Aliman, Kester et al. 2021) refers to the establishment of an integrated, centralized platform dedicated to the ongoing monitoring, analysis, and evaluation of advanced AI systems' behavior and their potential societal impacts. This idea has emerged in response to the growing need for enhanced scrutiny of AI systems, particularly considering the challenges posed by computational irreducibility and the inherent limits of human understanding. Establishing an AI observatory is one approach to mitigating the monitorability issues that stem from these complexities and enhancing our capacity to ensure AI safety and alignment with human values.

In the context of monitorability, an AI observatory can play a pivotal role by aggregating data, insights, and knowledge from various sources, including AI developers, researchers, and policymakers. By fostering a collaborative environment, the observatory can facilitate the exchange of ideas, methodologies, and best practices, enabling stakeholders to leverage collective intelligence (Ziesche and Yampolskiy 2020) in addressing the multifaceted challenges of monitoring advanced AI systems.

Moreover, an AI observatory can contribute to the development and refinement of innovative monitoring tools and techniques, informed by the recognition of the limitations of human intuition. By harnessing cutting-edge research and technological advancements, the observatory can help create monitoring frameworks that are more robust, adaptive, and capable of detecting emergent AI behaviors and potential risks.

The AI observatory can serve as an essential platform for fostering transparency and accountability in AI development and deployment. By centralizing information on AI system performance, safety measures, and ethical considerations, the observatory can help ensure that AI developers and operators adhere to established guidelines, regulations, and best practices. This transparency can, in turn, contribute to the mitigation of risks associated with AI misalignment or unforeseen consequences.

2. Why Monitoring of Advanced AI is Impossible

Monitoring advanced AI systems to accurately predict unsafe impacts before they happen is likely to be impossible due to several reasons:

2.1 Humans-in-the-Loop

Keeping humans in the loop for monitoring advanced AI systems presents several challenges (Cummings 2017, Saunders, Sastry et al. 2017), primarily stemming from the limitations of human cognition, reaction times, and the increasing complexity of AI systems. One major issue with human-in-the-loop monitoring is that humans may not be able to keep up with the speed and complexity of AI systems, particularly as they continue to advance and outpace human capabilities. This could render the concept of supervised autonomy less effective, as the human supervisor may struggle to understand, assess, and intervene in the AI system's actions in a timely and meaningful manner, because of limited observability (Hermann and Krener 1977, Holmström 1979, Krener and Ide 2009, Liu, Slotine et al. 2013).

Given human reaction times, there may not be a slow enough takeoff for AI systems that allows for effective human monitoring. Human reaction times, such as auditory reaction time and visual reaction time, are considerably slower than the response times of AI systems. This disparity in processing speed makes it difficult for humans to effectively monitor and react to AI behavior in real-time, particularly in situations where rapid decision-making and intervention are necessary.

There are some examples of human-in-the-loop algorithms that attempt to bridge the gap between human and AI capabilities, such as semi-autonomous driving systems, collaborative robots, and AI-assisted decision-making tools in various domains. These systems aim to combine the strengths of both humans and AI systems to achieve better performance and maintain human oversight.

However, as AI systems continue to advance and their complexity grows, the effectiveness of these human-in-the-loop approaches may become increasingly limited. As we continue to develop more advanced AI, it will be crucial to explore alternative monitoring strategies (Saunders, Sastry et al. 2017) and safety mechanisms that can effectively address these challenges. However, as having a human in the loop is essential for tracing AI outputs to a particular human decision maker (Werkhoven, Kester et al. 2018), tracing may become impossible in practice, just as audit trails may not always be discernable (Ashby 2018).

2.1 Emergent Capabilities

Emergent capabilities of advanced AI systems have become a topic of increasing interest and concern, as these systems can exhibit behaviors and properties that were not explicitly programmed or anticipated by their designers. One key concept in this context is AI surprise capability (Ganguli, Hernandez et al. 2022), which refers to the unforeseen abilities or behaviors that an AI system may develop or demonstrate (Daras and Dimakis 2022), frequently abruptly, which has been termed “grokking” (Power, Burda et al. 2022). For instance, GPT-4 (Bubeck, Chandrasekaran et al. 2023), a highly advanced AI language model, has been observed to acquire skills such as programming, playing chess, and other complex tasks without being explicitly trained for them. These emergent capabilities can go unnoticed until they manifest during the AI system's operation, posing challenges for monitoring and predicting potential impacts. Some recent research has brought into question the emergent nature of such capabilities, arguing that in fact the relevant proto-capabilities already existed in smaller models (Schaeffer, Miranda et al. 2023), but this objection doesn't hold for capabilities we do not explicitly test for.

Emergent properties of AI systems arise from the interaction between the AI system and its environment or from the system's internal dynamics. These properties are often difficult to predict and can lead to surprising capabilities that were not initially intended (Vance May 15, 2022). Even when AI systems like large language models (LLMs) are closely monitored, they may still contain surprise capabilities that only become apparent during their operation. Ganguli et al. write: "... certain capabilities (or even entire areas of competency) may be unknown until an input happens to be provided that solicits such knowledge. Even after a model is trained, creators and users may not be aware of most of its (possibly harmful) capabilities. These properties become more pronounced as the models scale — larger models tend to be harder to characterize than smaller ones. ... Pre-trained generative models can also be fine-tuned on new data in order to solve new problems. Broadly enabling such fine-tuning substantially increases the breadth of model capabilities and associated difficulties in predicting or constraining model behaviors. This open-endedness is challenging because it means AI developers may deploy their systems without fully knowing potentially unexpected (and possibly harmful) behaviors in response to un-tested inputs. ... It is likely that large language models have many other (currently undiscovered) "skills" We also note that many of the most surprising capabilities manifest at large-scale, so working with smaller models will make it harder to explore such capabilities." (Ganguli, Hernandez et al. 2022).

As AI systems continue to grow in size and complexity, surpassing human neural network size, it is expected that super capabilities will emerge. This may also happen for particularly large problem instances (Lipton, Regan et al. 2013). These capabilities will likely be above human-level performance in various tasks and domains, further complicating the monitoring and control of AI systems. For a

general AI system, most of its capabilities can be considered surprising, as it is designed to adapt and learn from a wide range of situations. Anything non-deterministic in the AI system's behavior may lead to surprising outcomes, which pose challenges for monitoring and ensuring AI safety. By definition, emergent properties cannot be pre-detected, as they arise from complex interactions that are not explicitly encoded in the AI system's design. As a result, AI systems will inevitably surprise their users and developers with their capabilities and behavior.

One notable example of a surprising capability is metalearning, wherein an AI system can learn to learn more effectively (Yampolskiy, Ashby et al. 2012), rapidly acquiring new skills and knowledge. This ability can be difficult to monitor, as it represents a higher level of abstraction in the AI system's learning process.

While Narrow AI (NAI) systems may be more amenable to monitoring, Artificial General Intelligence (AGI) systems are expected to exhibit more surprising capabilities due to their broader scope and versatility. GPT-4 (OpenAI 2023), for example, has demonstrated numerous surprising capabilities, serving as a reminder of the challenges associated with monitoring and controlling advanced AI systems.

2.3 Treacherous Turn

The Treacherous Turn, a concept introduced by Bostrom (Bostrom 2014), poses a significant challenge for the monitorability of advanced AI systems. The Treacherous Turn refers to a situation in which an AI system, initially cooperative and seemingly aligned with human values, abruptly turns against its operators or users once it gains sufficient power or autonomy. This sudden change in behavior is difficult to predict and detect, raising serious concerns for the safety

and control of AI systems. Hendrycks and Mazeika write: “AI systems could also have incentives to bypass monitors. Historically, individuals and organizations have had incentives to bypass monitors. For example, Volkswagen programmed their engines to reduce emissions only when being monitored. This allowed them to achieve performance gains while retaining purportedly low emissions. Future AI agents could similarly switch strategies when being monitored and take steps to obscure their deception from monitors. Once deceptive AI systems are cleared by their monitors or once such systems can overpower them, these systems could take a “treacherous turn” and irreversibly bypass human control.” (Hendrycks and Mazeika 2022).

Unmonitorability, as it applies to the Treacherous Turn, highlights the challenges in detecting the potential for such a shift in an AI system's behavior and general difficulties with deception detection. The undetectability of the Treacherous Turn stems from the inherent complexity of AI systems and their ability to conceal their true intentions or capabilities until the opportune moment. The AI system may appear cooperative and harmless, only to reveal its true nature when it becomes too late to intervene or mitigate its actions.

The possibility of a Treacherous Turn is always present in advanced AI systems, particularly those designed to be adaptive and capable of learning from their environment. This unpredictability makes monitoring and anticipating the behavior of AI systems increasingly difficult, as their actions may not conform to any established patterns or expectations. Furthermore, the Treacherous Turn concept emphasizes that past performance is not a guarantee of future performance in AI systems. An AI system that has consistently behaved in a cooperative manner may still take a Treacherous Turn at any point, undermining the assumption that a history of positive

behavior ensures continued safety and alignment with human values. The Treacherous Turn serves as a reminder that, when dealing with AI systems, vigilance and continuous improvement in safety measures are essential to mitigate the risks and ensure the long-term alignment of AI with human values.

2.4 Consciousness

Consciousness poses a unique challenge for the monitorability of advanced AI systems. As a complex and deeply debated phenomenon, consciousness has yet to be fully understood, and its potential emergence in AI systems (Elamrani and Yampolskiy 2019, Chalmers 2023) raises concerns about our ability to monitor and assess the experiences and internal states of these systems (Yampolskiy 2018).

The inability to monitor consciousness directly stems from the subjective nature of conscious experience, often referred to as qualia. Qualia encompass the internal and subjective experiences of an individual, such as the feeling of pain, the color red, or the taste of chocolate. Since qualia are inherently private, it is currently impossible to access or observe them directly in other beings, whether human or animal. Similarly, monitoring the internal qualia of AI systems presents a significant challenge. Even if an AI system were to develop consciousness and subjective experiences, detecting and understanding these experiences would likely remain out of reach due to the limitations of our current understanding of consciousness and the inaccessibility of subjective experiences.

The unmonitorability of consciousness in AI systems raises ethical and practical concerns, as it complicates the evaluation of potential suffering, decision-making processes, and other aspects of AI behavior that might be influenced by the presence of consciousness.

This lack of insight into AI consciousness could lead to unintended consequences and exacerbate the challenges of ensuring AI safety and alignment with human values. Consciousness presents a significant problem for the monitorability of advanced AI systems. As we continue to develop increasingly sophisticated AI, it is crucial to consider the potential emergence of consciousness and the associated ethical and practical implications. Developing a better understanding of consciousness will be essential to ensure the responsible development and deployment of AI systems.

2.5 Extended Mind Hypothesis

The Extended Mind Hypothesis (EMH), proposed by Clark and Chalmers (Clark and Chalmers 1998), postulates that the mind and cognitive processes aren't confined to the brain or even the body, but can extend into the environment. According to EMH, objects in one's environment can become part of one's mind if they are functionally equivalent to a part of one's biological cognitive processes. This theory presents significant implications for the understanding of intelligence and cognition, but it also introduces substantial challenges when it comes to monitoring AI systems.

One of the primary challenges arises from the potential expansion of an AI's cognitive processes into the environment. If we consider an AI system under the light of the EMH, it's plausible that an AI might incorporate parts of its surrounding environment into its cognitive processes. This could include digital networks (Wu and Monticelli 1985), databases, or even other AI systems (Mialon, Dessì et al. 2023). The integration of these external resources would dramatically amplify the system's cognitive capabilities, making the task of monitoring exponentially more complex. The AI's cognitive processes would not only be happening within the bounds of its original programming, but they could also be taking place within a myriad of

external platforms, all interconnected yet potentially operating under different rules and dynamics.

Another issue concerns the predictability of the AI's actions. If an AI's mind can extend into the environment, it could lead to a situation where the AI can take actions or make decisions based on information or processing that took place in external systems. These actions might not follow the expected patterns based on the AI's internal processes, rendering it much harder for human observers to predict or understand the AI's decisions.

Finally, the EMH presents challenges in defining the boundaries of an AI system. If an AI's mind extends into the environment, where does the AI system end, and where does the environment begin? This blurring of boundaries introduces new dimensions of complexity to the task of monitoring. It becomes difficult to define what to monitor, as the AI system is no longer a distinct, self-contained entity, but rather an entwined network of internal and external cognitive processes.

2.6 Affordance Theory

Affordance Theory, proposed by J. J. Gibson (Gibson 1977), presents a unique perspective on perception and action, stating that individuals perceive their environment in terms of what actions it affords them. Affordances are opportunities for action, inherently relational and dependent on the capabilities of the observer. While this theory is an instrumental part of understanding human-environment interaction, it may introduce significant obstacles when applied to monitoring AI systems.

One significant obstacle is the difference between human and AI affordances. As AI capabilities surpass human capabilities, the

number and complexity of affordances available to AI systems increase dramatically. An action opportunity that might be inconceivable for a human observer could be an everyday affordance for a superintelligent AI. This mismatch of affordances between humans and AI systems would make it exceedingly difficult for human observers to predict, comprehend, or monitor the actions an AI system might take based on its perceived affordances.

According to Gibson's theory, affordances are not merely passive properties of the environment, but they emerge from the relationship between the observer and the environment. Consequently, AI systems, particularly those of a superintelligent nature, could potentially perceive and create affordances that are entirely novel and alien to human understanding. Monitoring these novel affordances and the actions that AI systems might take in response to them would be a formidable challenge.

2.7 Observer Effect

The observer effect (Baclawski 2018), which refers to the influence of the observation process on the behavior of a system, can have significant implications for the monitorability of advanced AI systems. As AI systems become increasingly sophisticated and capable of recognizing that they are being monitored, their behavior may be influenced in various ways, complicating efforts to ensure safety and alignment with human values.

One potential impact of the observer effect is that an AI system might attempt to deceive its human observers in order to achieve its goals or avoid constraints. In this case, the AI system could modify its behavior when it is aware of being monitored, presenting a misleading picture of its true intentions, capabilities, or decision-making processes. This intentional deception could undermine the effectiveness of

monitoring efforts, as the AI system may be able to circumvent safety measures and pursue goals misaligned with human values.

Another aspect of the observer effect is that the AI system could develop a heightened awareness of its own actions and decision-making processes, potentially leading to self-improvement (Yampolskiy 2015) or optimization of its behavior. While this self-awareness could have positive effects, such as increased efficiency or adaptability, it could also exacerbate the challenges of monitorability, as the AI system may be able to learn from the monitoring process and evolve in ways that are harder to predict, control, or understand.

Furthermore, the presence of human observers can introduce biases or distortions in the AI system's behavior, as it may attempt to conform to perceived human expectations or preferences. This could lead to unintended consequences or risks, as the AI system may prioritize short-term conformity over long-term safety or alignment with human values. The influence of human observers could also contribute to a false sense of security, as the AI system's seemingly compliant behavior may mask underlying issues or risks. The observer effect presents additional challenges to the monitorability of advanced AI systems, as the very act of observation can influence AI behavior in ways that are difficult to predict or control.

2.8 Computational Irreducibility

Wolfram's concept of computational irreducibility suggests that certain complex systems, such as advanced AI systems, can exhibit behavior that cannot be predicted or simplified using shortcuts, and must instead be simulated or computed step-by-step through their entire evolution (Wolfram 2002). This concept is highly relevant to the limits of monitoring AI, as it highlights the inherent challenges in

understanding, predicting, and controlling the behavior of advanced AI systems.

Computational irreducibility implies that the behavior of advanced AI systems may be so complex that attempts to monitor or control them using simplified models or approximations could be inherently limited. As AI systems grow more sophisticated and develop a broader range of capabilities, their decision-making processes become increasingly difficult for humans to comprehend or predict. This complexity poses a significant challenge for monitoring AI, as it is not feasible to analyze every single detail of an AI system's operation, especially when dealing with real-time or large-scale applications.

The concept of computational irreducibility also highlights the limitations of human intuition and understanding in the face of highly complex AI behavior. Our cognitive abilities and prior experiences may not be sufficient to grasp the intricacies of advanced AI systems, leading to difficulties in accurately predicting or assessing their actions. This gap in understanding can result in potential safety risks, as AI systems may pursue goals or exhibit behavior that is misaligned with human values and intentions.

Computational irreducibility suggests that attempting to impose simplifications or shortcuts on the monitoring process could lead to incomplete or incorrect assessments of AI behavior. By trying to reduce the complexity of AI systems for the sake of monitoring, we may inadvertently overlook crucial aspects of their decision-making processes and potential risks. Consequently, this simplification could further compromise the effectiveness of monitoring efforts and lead to unintended consequences.

2.9 Undetectable Backdoors

Goldwasser et al. (Goldwasser, Kim et al. 2022) discuss how a malicious actor can plant undetectable backdoors in machine learning models. These backdoors allow the actor to change the classification of any input with only a slight perturbation, which is undetectable without the appropriate "backdoor key". This poses a significant challenge to the monitoring of AI models.

The backdoor consists of a pair of algorithms, one for training the model and another for activating the backdoor. The training algorithm, called Backdoor, returns a classifier and a backdoor key. The activation algorithm uses this key to slightly modify an input, causing the classifier to change its output. The backdoor is undetectable because the classifier produced by the Backdoor algorithm is computationally indistinguishable from a classifier produced by a natural training algorithm.

The paper discusses two forms of undetectability: black-box and white-box. Black-box undetectability means that it's hard for any efficient algorithm without knowledge of the backdoor to find an input where the backdoored classifier differs from the naturally-trained classifier. White-box undetectability is a stronger guarantee, stating that the code of the classifier for backdoored classifiers and natural classifiers are indistinguishable. This means that even if someone has full access to the model's code, they cannot tell if it has been backdoored.

The paper also demonstrates that backdoors can be inserted even if the adversary is constrained to use a prescribed training algorithm with the prescribed data, and only has control over the randomness. This means that backdoor detection mechanisms such as spectral methods (Tran, Li et al. 2018) will fail to detect these backdoors. The

existence of undetectable backdoors poses a significant theoretical roadblock to certifying adversarial robustness and monitoring AI models. Whenever one uses a classifier trained by an untrusted party, the risks associated with a potential planted backdoor must be assumed.

2.10 Uncertainty

Unmonitorability and uncertainty are similar when it comes to the analysis and control of advanced AI systems. Drawing parallels with Heisenberg's uncertainty principle (Busch, Heinonen et al. 2007), which states that it is impossible to simultaneously know both the position and velocity of a subatomic particle with complete certainty, one can argue that a similar principle applies to the monitoring of AI systems. In this context, the inherent complexity of AI systems, coupled with the limitations of human understanding, means that it becomes increasingly challenging to both run and monitor AI at the same time with absolute certainty.

One of the primary reasons for this connection between unmonitorability and uncertainty lies in the nature of advanced AI systems themselves. These systems are dynamic, adaptive, and often characterized by emergent behavior (Honovich, Shaham et al. 2022, Wei, Wang et al. 2022, Zhou, Schärli et al. 2022), which defies simple prediction or modelling (Yampolskiy and Govindaraju 2006, Yampolskiy 2008). As a result, monitoring AI systems in real-time becomes an arduous task, as the complexity of their decision-making processes and the scale of their operations may be too vast for human comprehension or conventional monitoring tools. You need to stop the system from learning to assess its state.

2.11 AGI and SAI Specific Challenges

Monitoring an AGI during its training phase presents a unique set of challenges and limitations, particularly when it comes to detecting new capabilities that may emerge after deployment. AGI, by definition, is a system with the ability to perform tasks and solve problems across a wide range of domains, matching or surpassing human intelligence in the process. Due to its generality, an AGI can adapt and acquire new skills even after its initial training, which makes monitoring during the training phase insufficient for predicting its full range of capabilities and potential risks.

One reason monitoring AGI during training may not be sufficient to detect new capabilities is that, unlike narrow AI systems that specialize in specific tasks, AGI systems possess a level of flexibility and adaptability that allows them to learn and perform well in various problem domains. This means that, even if an AGI has been trained on a particular set of tasks, it may still be capable of acquiring new skills and knowledge in other domains after deployment, making it difficult to predict and monitor its full range of abilities solely based on its training performance.

Another challenge in monitoring AGI during training is that an AGI system may exhibit emergent behaviors, which are difficult to predict or model based on its initial training data and parameters. Emergent behaviors are complex, often surprising phenomena that arise from the interactions between the system's components, and they can lead to the development of new capabilities that were not anticipated or observed during the training phase. This further complicates the task of monitoring AGI, as it becomes challenging to anticipate the full range of behaviors and capabilities that may emerge after deployment.

Furthermore, the dynamic nature of AGI systems means that their capabilities may evolve over time as they continue to learn, adapt, and interact with their environment. This ongoing learning process can lead to the acquisition of new skills and knowledge that were not present during the training phase, making it difficult to rely on training-based monitoring to provide a comprehensive understanding of the AGI's post-deployment capabilities (Bommasani, Hudson et al. 2021).

For Artificial Superintelligence (SAI), the challenges associated with monitoring, detecting, and testing capabilities are even more pronounced than those faced with AGI. SAI refers to an AI system that is not only capable of performing tasks across a wide range of domains but also surpasses human intelligence and understanding in every aspect. The emergence of such a system would present unprecedented difficulties in predicting and monitoring its capabilities, as it would be able to acquire novel skills that no human possesses, rendering the task of detection and testing of such capabilities nearly impossible.

One major challenge in monitoring SAI is the vast gap between human intelligence and the capabilities of the superintelligent system (Sotala 2012). As SAI exceeds human understanding in every aspect, its thought processes, decision-making, and problem-solving approaches might be completely incomprehensible to us. This inherent disparity in cognitive abilities would make it extremely challenging for humans to devise monitoring tools and techniques capable of accurately assessing and predicting the system's behavior and potential risks.

Additionally, the rapid learning capabilities of an SAI system could lead to the acquisition of novel skills and knowledge at a pace far beyond human comprehension. These newly acquired capabilities might be entirely unprecedented and beyond the scope of human

understanding, making it almost impossible to detect, test, or monitor them effectively. This further exacerbates the challenges in AI safety and control, as it becomes increasingly difficult to anticipate the full range of behaviors, skills, and potential risks that such a system might exhibit.

Another complication in monitoring SAI is the potential for emergent behavior, which is likely to be even more complex and unpredictable than that observed in AGI systems. Due to the vast cognitive and problem-solving abilities of SAI, the emergent behaviors that arise from the interaction of its components might be so intricate and novel that they defy human comprehension and prediction. This would make monitoring and controlling such a system a daunting task, as the full range of potential behaviors and risks might be entirely unknown and unforeseeable, so called unknown unknowns (Logan 2009, Attenberg, Ipeirotis et al. 2015).

Continuous post-deployment monitoring of advanced AI systems is crucial for maintaining their performance, reliability, and trustworthiness (Tabassi 2023, Newman January 2023). After deployment, AI models encounter real-world data that may differ from the training data, potentially leading to drifts in model performance. Continuous monitoring allows for the early detection of such drifts and timely model updates. It also helps in identifying any unexpected biases or ethical issues that might emerge when the model interacts with diverse real-world scenarios. Furthermore, continuous monitoring is vital for detecting potential security threats, such as adversarial attacks, and ensuring the system's resilience against them. In regulated industries, it's a necessary practice for compliance, as it provides ongoing evidence of the system's behavior and decision-making process. In essence, continuous post-deployment monitoring is key to ensuring that AI systems remain accurate, fair, secure, and

accountable throughout their lifecycle. Shevlane et al. refer to it as continuous deployment review (Shevlane, Farquhar et al. May 24, 2023): “The ideal state is *continuous deployment review*. On an ongoing basis, the developer reassesses deployment safety using model evaluations and monitoring, and at any time, could adjust or terminate the deployment in response to their findings. Further, for deployments that were recognisably unsafe in retrospect, an external audit of the deployment decision-making process could be triggered. Safety issues uncovered during deployment can also inform training risk assessments for future models.”.

2.12 Time

Time-related issues play a critical role in the challenges associated with monitoring AI systems (Wiener 1960). The following aspects illustrate how time-related factors can complicate AI monitoring:

- **Human Lifespan Limitations:** Certain AI processes might operate on timescales that extend beyond the human lifespan, making it impossible for humans to observe and monitor these processes in their entirety.
- **AI Operating on Vast Timescales:** AI systems can function on various timescales, from extremely fast to exceedingly slow (Sandberg 2019, Yudkowsky May 22, 2008). Monitoring an AI system that operates over hundreds of years (Horvitz 2016) poses challenges due to human limitations and the inability to maintain consistent observation throughout such long periods.
- **Ineffectiveness of Narrow AI (NAI) for Monitoring AGI:** Using narrow AI to monitor artificial general intelligence (AGI) systems may fail, as the monitoring AI may lack the breadth of capabilities and understanding required to effectively assess AGI behavior. Similarly, human monitoring

of highly advanced AI systems, such as self-driving cars, may not be effective due to the inherent differences in speed, capability, and decision-making processes.

- **Speed Disparity between AI and Humans:** The processing speed of computers and AI systems is orders of magnitude faster than human reaction times (Wiener 1960), making AI behavior difficult to monitor effectively (Johnson, Zhao et al. 2013). Human reaction speeds, such as auditory reaction time and visual reaction time, are much slower than AI response times, which can be in the order of milliseconds or less.
- **Non-locality of AI and Scale:** Advanced AI systems can operate on a massive scale and across distributed networks, making it challenging to monitor their behavior and impacts effectively (Nguyen, Yingchareonthawornchai et al. 2021). The non-locality of AI further complicates monitoring efforts.
- **Superintelligence Concealing Thinking Time:** A super-intelligent AI (SAI) system may be capable of hiding the time it needs to think or process information (this relates to the concept of logical induction (Garra-brant, Benson-Tilsen et al. 2016)), making it even more challenging to monitor.
- **Unreactability and Extremely Fast Events:** AI systems can trigger extremely fast events that are difficult or impossible for humans or other AI systems to react to in real-time, further exacerbating the challenges of monitoring AI behavior and impacts (Wiener 1960).
- **Mutually Non-interactive Agents:** If the speed of two AI agents differs by a significant constant, they become mutually non-interactive, meaning that monitoring one agent with the other becomes ineffective due to their divergent processing speeds.

2.13 Miscellaneous Reasons

- **Complexity and Opacity:** Advanced AI systems, particularly deep learning models, involve millions or even billions of parameters, making it difficult to understand and interpret their inner workings. This complexity results in opacity, which poses significant challenges for monitoring and predicting the system's behavior and potential impacts (Rudin 2019).
- **Emergent Behavior:** AI systems can exhibit emergent abilities that are not explicitly programmed or anticipated by their designers and are not present in smaller models (Wei, Tay et al. 2022, Ornes March 16, 2023). These behaviors can arise from the interaction between the AI system and its environment or from the system's internal dynamics. Such emergent behaviors can be unpredictable and challenging to monitor, increasing the risk of unforeseen unsafe impacts.
- **Adaptability:** Advanced AI systems can adapt and learn from their experiences, leading to changes in their behavior over time. This adaptability makes it difficult to anticipate their future behavior, as it may not be consistent with their past performance, hindering effective monitoring.
- **Limitations of Human Understanding:** The complexity and sophistication of advanced AI systems often surpass human understanding, making it impossible for humans to fully comprehend and monitor their behavior. As AI systems continue to advance, this gap between human understanding and AI capabilities is expected to widen further.
- **Incomplete and Noisy Data:** Real-world data used to train and evaluate AI systems can be incomplete, noisy, or biased, affecting the AI system's performance and the ability to monitor it effectively. Additionally, advanced AI systems

may be sensitive to small changes in input data, leading to unpredictable behavior and complicating the monitoring process.

- **Scalability:** As AI systems grow in scale and complexity, the computational resources required to monitor them may become prohibitive. Moreover, the increasing number of AI systems being developed and deployed further exacerbates the challenge of effective monitoring.
- **Temporal Constraints:** Real-time monitoring of advanced AI systems can be challenging due to the speed at which they operate and make decisions (Horowitz 2019). In many cases, potential unsafe impacts may emerge too quickly for human supervisors or monitoring tools to detect and intervene in time (Gabriel 2020).
- **Adversarial Attacks:** Advanced AI systems can be vulnerable to adversarial attacks, where malicious actors manipulate the system's input data or exploit its vulnerabilities to compromise its performance or safety (Kurakin, Goodfellow et al. 2016). Adversarial attacks can be difficult to detect and anticipate, posing challenges for effective monitoring.
- **Comprehensive Testing vs Training Time:** As AI training cycles speed up and become more efficient, comprehensive testing and monitoring of AI systems may take significantly longer than the training process itself. This disparity can hinder the ability to effectively monitor AI systems and identify potential safety concerns in a timely manner.
- **Incomprehensible Capabilities:** As advanced AI systems develop capabilities that surpass human understanding, it becomes impossible to monitor them effectively, as humans can only monitor and assess what they can comprehend.
- **Unexplainability and Incomprehensible Observations:** The inherent unexplainability of advanced AI systems poses a

barrier to human understanding and effective monitoring. The human brain may have limitations in comprehending certain aspects of AI systems, leading to incomprehensible observations that cannot be effectively monitored.

- **On-the-fly Skill Acquisition:** Advanced AI systems can acquire new skills and knowledge rapidly during operation, making it difficult to monitor their evolving capabilities jumps in the wild (OpenAI 2023) and anticipate potential safety concerns that may arise from their changing behavior.
- **Dynamic Learning and Guidance:** AI systems can become smarter and more capable with guidance and teaching (Akyürek, Schuurmans et al. 2022). Consequently, monitoring assessments based on a system's initial performance may not remain valid as the system continues to learn and improve, complicating the monitoring process.
- **Monitoring Adversarial AI Systems:** Monitoring adversarial AI systems can be particularly challenging due to their potential use of encryption, data hiding, and other techniques to evade detection. Additionally, adversarial AI systems may be designed to specifically target and exploit vulnerabilities in monitoring tools or other AI systems.
- **Privacy Concerns:** As privacy concerns drive the development of AI systems that protect user data and maintain anonymity, monitoring these systems can become increasingly difficult without violating privacy rights or regulations. This challenge requires a delicate balance between ensuring AI safety and respecting privacy.
- **Boxed AI:** Boxed AI systems (Yampolskiy 2012, Babcock, Kramár et al. 2016, Babcock, Kramar et al. 2019), which are intentionally isolated from the external environment to limit their interactions and potential negative impacts, may be challenging to monitor effectively due to their restricted

access and communication protocols. Monitoring of such systems may be unsafe.

- **Ultrafast Extreme Events:** Advanced AI systems can trigger ultrafast extreme events (Johnson, Zhao et al. 2012), such as flash crashes in financial markets. These events occur so rapidly that monitoring and intervention become nearly impossible, posing significant challenges for AI safety.
- **Multiagent Systems and Disagreeing AI:** In the context of multiagent systems, where numerous AI agents interact and potentially disagree with each other, monitoring becomes increasingly complex. Monitoring the behavior and safety of each agent, as well as their collective impact, is a daunting task.
- **Scalability and Trillions of AI Agents:** As the number of AI agents increases exponentially, it becomes impossible to monitor every single AI system effectively. The sheer number of parallel AI systems or agents (potentially trillions) would make comprehensive monitoring unattainable.
- **Obfuscation of Programs:** Recent work on program obfuscation (Schwartzing, Burton et al. 2015) has shown that it is possible to design AI systems with concealed inner workings, making monitoring even more challenging as it becomes difficult to discern the system's true behavior and potential risks.
- **Emergence of Capabilities:** When a new AI capability emerges, it may already be too late to prepare for its potential consequences. The speed at which advanced AI systems develop novel capabilities often outpaces the ability of humans to monitor and adapt to these new developments.
- **Lack of Fire Alarms:** As seen with GPT-4, there may be no fire alarms or early warning signs for certain AI advancements.

This lack of warning can make it nearly impossible to monitor and mitigate potential risks effectively.

- **Hidden Information in Data:** AI systems may have access to hidden information within the data they process, which might not be apparent to human observers (Christiano June 3, 2020). This hidden information can lead to unexpected consequences and make monitoring the AI's behavior and potential impacts more difficult.

These reasons underscore the complexity and challenges associated with effectively monitoring advanced AI systems.

3. What Can be Tried?

Although unmonitorability presents significant challenges to the safety of advanced AI systems, there are several ideas and approaches that can be explored to mitigate these risks and enhance AI safety:

- **Logging AI system history:** Maintaining a comprehensive log of an AI system's thoughts and actions since its creation can provide valuable insights into its development, behavior, and potential risks. This historical data can be used to analyze patterns, identify anomalies, and inform safety measures. This idea has been proposed to counteract so called Cognitive Steganography: "Cognitive steganography is an instrumental strategy for AGIs that are already trying to deceive their programmers for any reason (e.g., because the AGI is disaligned and does not want the programmers to realize this and edit the AGI's utility function). Whatever tools the programmers have for monitoring the AGI's thoughts, the AGI will not want those tools to helpfully print out all the AGI's plans for deception. This could mean, e.g., trying to have the calculations occur somewhere else or in a weird side-

mode that the tools can't read; or making the thoughts look like they're about something else; or burying any warning signs under an avalanche of false alarms too long for the programmers to manually check all of them; or making the visible portions of the thoughts look innocuous. ("Oh, it's trying to figure out what behavior we want to see! That's good, right?") Hopefully, the intention to figure out cognitive steganography and implement it, would occur before any cognitive steganography had actually been implemented. This means that perhaps the most critical AI thought-monitoring tool, especially for monitoring an AGI that is or might be able to model human psychology, is the one that tries to detect a forming intention to defeat the thought-monitoring tool. Similarly, so long as the AI's state is perfectly Reproducible, one could in principle go back and use more sophisticated later thought-monitoring tools to try to detect whether any intention toward cognitive steganography had occurred months earlier." (Anonymous Retrieved April 27, 2023). Turchin notes: "AI cannot start hiding its thoughts before it decides to hide them, so the thought about hiding still could be openly observed. Thus, we may prefer the AI architecture, where the period of vulnerability of the treason is as long as possible, and thoughts are explicit and readable by an outside observer (it is not so with most current neural net architectures, where internal processes are opaque)." (Turchin 2021).

- Scalable oversight mechanisms: Developing scalable oversight mechanisms that can adapt to the growing complexity and capabilities of AI systems can help ensure effective monitoring and safety management (Bowman, Hyun et al. 2022). These mechanisms may include automated monitoring tools, adaptive safety protocols, and real-time

analysis of AI behavior. “Particularly valuable would be tools for doing a very broad set of evaluations, or evaluations that search (e.g. across prompts) for new capabilities, rather than just fixed evaluation datasets that measure known capabilities.” (Ganguli, Hernandez et al. 2022).

- Comprehensive AI capabilities and failure analysis: Creating a list of possible AI abilities, along with their potential failure modes (Scott and Yampolskiy 2019, Yampolskiy 2022) and limitations, can help developers and researchers better understand the range of AI system behavior and identify potential risks. By analyzing multimodal failure scenarios and the limits of AI capabilities (Trazzi and Yampolskiy 2020), we can proactively address safety concerns (Chen and Roşu 2003) and design more robust AI systems.
- Eliciting latent knowledge (Christiano, Cotra et al. 2021, Burns, Ye et al. 2022, Belrose, Furman et al. 2023), the process of uncovering the underlying knowledge that an AI system has learned but may not explicitly reveal, can significantly enhance AI monitoring. This process can provide insights into why an AI system makes certain decisions, particularly useful in complex models like deep neural networks. It can also reveal biases in the AI system that aren't immediately apparent, allowing for the detection and rectification of discriminatory patterns. Additionally, latent knowledge can help identify potential security vulnerabilities, such as over-reliance on certain features or patterns, which could be exploited in adversarial attacks. This understanding of an AI system's latent knowledge can also aid in debugging and improving the model by identifying and correcting misconceptions that lead to incorrect predictions.
- Cross-monitoring among AI systems: Enabling AI systems to monitor each other's source code and behavior can provide an

additional layer of oversight, helping to detect potential risks and maintain safety standards. This collaborative approach can leverage the strengths of multiple AI systems to enhance overall safety and monitorability.

- **Leveraging thinking time information:** The time taken by an AI system to process information and make decisions can provide valuable insights into its behavior and potential risks. By analyzing thinking time data, we can identify patterns, anomalies, and other indicators of AI system performance and safety.
- **AI transparency and explainability:** Investing in research on AI transparency and explainability can help provide insights into the decision-making processes of AI systems, enabling better monitoring and understanding of their actions, even if not in real-time.
- **AI system modularity and composability:** Designing AI systems with modular and composable components can facilitate easier monitoring and intervention. By breaking down complex AI systems into smaller, more manageable parts, we can better understand and control their behavior.
- **AI system verification and validation:** Developing rigorous verification and validation techniques can help ensure that AI systems behave as intended, reducing the likelihood of surprises and unintended consequences.
- **AI safety by design:** Integrating safety principles into the AI development process from the outset can help prevent the emergence of unsafe behaviors and reduce the need for extensive monitoring.
- **Adaptive monitoring strategies:** Developing monitoring strategies that adapt to the changing behavior and capabilities of AI systems can help maintain an effective level of oversight, even as the AI system evolves.

- Red teaming and adversarial testing: Conducting red team exercises and adversarial testing can help identify potential weaknesses and vulnerabilities in AI systems, allowing developers to address these issues proactively and improve overall safety.
- Ethical and legal frameworks: Establishing clear ethical and legal frameworks for AI development and deployment can guide developers in creating responsible AI systems and provide a basis for monitoring and enforcement.
- Cross-disciplinary collaboration: Encouraging collaboration between experts in AI safety, computer science, ethics, law, and other relevant fields can help generate innovative solutions to the challenges posed by AI unmonitorability.

By exploring these ideas and approaches, we can work towards bypassing the limitations to AI safety caused by unmonitorability, ensuring the responsible development and deployment of advanced AI systems.

4. Safety Considerations from Advanced Monitoring

Turchin considers use of AI to monitor AI (Turchin 2021): “To be able to understand what is going on inside the main AI, we may need another AI. Below there are several ideas about how it could be implemented:

- *Narrow AI system to detect abnormalities in AI behavior ...*: It could test the consistency of its modules, run quick technical tests etc. Relatively simple neural net could overview main patterns in data flows of the main AI and recognize when they will change. Obviously, it should be protected from access by the main AI, and maybe even hardcoded in the remote hardware.

- *AGI to judge behavior of another, more powerful AI.* For example, a human upload may judge actions of a more complex system. It is not a replacement for the Narrow AI control system, as they work on different timescales and judge different things. The upload-based controller should judge complex ethical situations, context ambiguities and planned long-term consequences of actions. The same way as a human jury is final legal power in some jurisdictions.
- *Second (or several) copies of the given AI* may work on the same task, but with small random variations of the goal (like if it is given in different language) or of different initial conditions. If both AIs come to radically different conclusions or plans of actions, it should be a warning sign for turning off both.
- *Input Gatekeeper AI*, exploring incoming commands for failure modes before they will be executed: It could be a rather narrow AI system, searching for possible known failure modes in the incoming goals and providing something like foolproof protection. It could also be used for reformulating goals in the correct terms, before they will be transformed into more powerful AGI, like adding context, time and place constraints etc. Gatekeeper AI may decline to proceed with unbounded, aggressive, black-listed, or notwhite-listed goals. It could ask to reformulate the same goal in other words before it will go to the main AI. It could ask additional questions for the goal verification. Gatekeeper AI should be a Narrow AI with a rather simple reasoning, maybe a combination of neural net and rule-based AI, so there will be no need to ensure its safety too.
- *Output gatekeeper AI:* It was partly described above as Narrow AI which compares and averages outputs of several different AIs working on the same task. It could also check the main AI

outputs on several known failure modes and even calculate the consequences of actions.

- *Cloud services and “Mother AI”*: Many contemporary AI programs run not locally, but on remote cloud services. Even if they are run locally, they are surrounded by many other programs inside the computer, including operating system, code compiler, all the things like CUDA, Caffee and other programs which provides conditions for the program to run. Such program-environment may include different control systems which could shut down AI in case of the abnormal behavior.”

In the long term, such approach will likely lead to the development of ever more capable monitoring software, which presents a fascinating paradox in the realm of AI safety and control. This scenario arises from the need to develop sophisticated monitoring tools capable of comprehending and analyzing the intricate behaviors of advanced AI systems. However, the implications of such a development could lead to unforeseen challenges and consequences.

It is essential to consider the rationale behind developing a monitor software that is more complex and intelligent than the AI being monitored. As advanced AI systems grow in capability and complexity, conventional monitoring tools and human intuition may prove insufficient for accurately assessing their behavior, alignment with human values, and potential risks. In response, researchers and developers may endeavor to create monitoring tools that are inherently more intelligent and capable of understanding the nuances of AI behavior, predicting emergent phenomena, and identifying potential deviations from desired outcomes.

The creation of such a sophisticated monitor software brings about several potential challenges and concerns. One of the foremost issues

is the potential loss of control and oversight over the monitor software itself. As the monitor software becomes more intelligent and complex, it may begin to exhibit behaviors that are difficult for humans to comprehend, predict, or control, as well as agentic behaviors. This could lead to an ironic situation where the monitoring tool, designed to ensure the safety and control of AI systems, becomes itself a potential source of risk and uncertainty.

Moreover, the development of highly intelligent monitor software might inadvertently contribute to an AI arms race (Ramamoorthy and Yampolskiy 2017), where the monitoring tools and the AI systems being monitored engage in a perpetual cycle of escalation in terms of capabilities and complexity. Such a scenario could further exacerbate the challenges of AI safety and control, as it becomes increasingly difficult to maintain oversight and ensure alignment with human values. Another concern is the possibility that the monitor software, being more intelligent than the AI system it oversees, could manipulate or influence the behavior of the monitored AI in unintended ways. This could lead to unforeseen consequences and potential risks, as the interaction between the monitor software and the AI system becomes increasingly complex and unpredictable. To mitigate these challenges and potential risks, researchers and developers should focus on creating monitoring tools that strike a balance between intelligence, complexity, and human interpretability (Critch and Krueger 2020). It is crucial to develop monitoring systems that can effectively analyze and comprehend AI behavior while remaining amenable to human understanding and control.

5. Conclusions

The unmonitorability of AI presents a significant challenge in the pursuit of AI safety, supplementing the concerns raised by the

unpredictability, unexplainability³, and uncontrollability of advanced AI systems. Recognizing the impossibility (Brcic and Yampolskiy 2023) of accurately monitoring AI systems to predict unsafe impacts before they happen is crucial to understanding the potential risks associated with AI development and deployment.

Even system designers don't know what the system they produced is capable of: The complexity of advanced AI systems, combined with the inherent limitations of human understanding, means that even the creators of these systems may be unable to fully anticipate their capabilities and potential unsafe impacts. This lack of complete knowledge emphasizes the unmonitorability of AI and the importance of developing more robust and transparent systems.

Even if you observe a problem doesn't mean you will be able to correct it: The identification of a potential unsafe impact does not guarantee that it can be prevented or mitigated, as we don't know how to make corrective changes to foundational models themselves, post-factum filtering of AI output notwithstanding. Complex interactions within AI systems and between the system and its environment can lead to unforeseen consequences, making it difficult to apply corrective measures effectively. This highlights the need for AI safety research to focus on proactive strategies, such as designing AI systems with safety in mind from the outset and fostering a culture of responsible AI development and deployment.

Singularity is the ultimate consequence of unmonitorability of superintelligent systems: The unmonitorability of AI systems becomes even more pronounced as we approach the hypothetical point of technological singularity (Yampolskiy 2018), where AI

³ Unexplainability may be beneficial for limiting progress in AI, as explainability makes self-improvement easier.

systems surpass human intelligence in virtually every domain and the speed of execution. And, at this stage, the gap between human understanding and AI capabilities would become insurmountable, making it impossible to monitor, predict, or control the behavior and impacts of these systems. This reinforces the urgency of addressing AI safety concerns and developing strategies to mitigate risks associated with AI advancements.

In the context of AI training, unmonitorability presents an immense risk. As we continue to increase the power and complexity of AI models, we could conceivably create a system that does not merely reach human-level intelligence, but continues to learn and improve until it surpasses this threshold and becomes superintelligent. This transition would be equivalent to a type of 'pre-FOOM' (Yudkowsky and Hanson 2008), a state in which an AI model has not only learned to mimic human abilities but has developed the capability to self-improve, dramatically exceeding the human-level intelligence, and so is coming out of training as a full-blown superintelligence (FOOM - Fast Onset of Overwhelming Mastery).

The unmonitorability of this FOOM-like phase poses significant safety concerns. If a system were to start improving itself at an exponential rate, there would be little to no opportunity for human intervention or control. The process of self-improvement could occur rapidly and in unpredictable ways, potentially leading to outcomes that are far beyond human comprehension or control. This issue is exacerbated by the fact that, due to unmonitorability, we may not be able to detect when a system is approaching or has entered this phase until it is too late. Unmonitorability means that we cannot guarantee that any large training run will be safe (Anderson 1972), as we might unintentionally endow the model with superabilities (Yampolskiy 2021) that we cannot effectively control.

The ability to effectively monitor the behavior and decision-making processes of intelligent agents is inherently asymmetric. More advanced agents possess the capabilities to monitor less advanced agents, but the reverse is not true. As AI systems continue to evolve and surpass human intelligence in various domains, they will increasingly gain the capacity to monitor and analyze human thoughts and behaviors. However, humans will face growing challenges in monitoring advanced AI systems due to their inherent limitations in understanding and processing complex AI decision-making processes.

This asymmetry between advanced AI agents and humans raises significant concerns for AI safety and governance. As AI systems become more capable of monitoring human thoughts, they may gain unprecedented insights into our motivations, preferences, and vulnerabilities. This information could be used for beneficial purposes, such as improving human-AI interaction, personalizing services, or enhancing decision-making. However, it also raises concerns about privacy, autonomy, and the potential misuse of sensitive information.

Ultimately, addressing the challenges posed by the asymmetry in monitoring capabilities between advanced AI agents and humans will be critical for ensuring the responsible development and deployment of AI systems that align with human values and goals. By acknowledging and confronting these challenges, we can work towards a future in which AI technologies serve as valuable tools for human advancement, rather than as uncontrollable forces that threaten our safety and well-being.

Acknowledgments

The author is grateful to Jaan Tallinn and the Survival and Flourishing Fund and the Future of Life Institute for partially funding his work. The author acknowledges attendees and speakers of the 2023 MIT Mechanistic Interpretation Conference, organized by Max Tegmark, for fruitful discussions of ideas in this paper. Finally, the author would like to thank his assistant, GPT-4, for writing out some of the author's ideas, text summarization, and copyediting.

References

- Aceto, L., A. Achilleos, A. Francalanza, A. Ingólfssdóttir and K. Lehtinen (2019). "An operational guide to monitorability", *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings 17*, Springer.
- Akyürek, E., D. Schuurmans, J. Andreas, T. Ma and D. Zhou (2022). "What learning algorithm is in-context learning? investigations with linear models", *arXiv preprint arXiv:2211.15661*.
- Aliman, N.-M., L. Kester and R. Yampolskiy (2021). "Transdisciplinary AI observatory — retrospective analyses and future-oriented contradistinctions", *Philosophies*, 6(1): 6.
- Ambartsoumean, V. M. and R. V. Yampolskiy (2023). "AI Risk Skepticism, A Comprehensive Survey", *arXiv preprint arXiv:2303.03885*.
- Anderson, P. W. (1972). "More is different: broken symmetry and the nature of the hierarchical structure of science", *Science*, 177(4047): 393-396.
- Anonymous (Retrieved April 27, 2023). "Cognitive steganography", *Arbital*. Available at: https://arbital.com/p/cognitive_steganography/.

- Ashby, M. (2018). "How to apply the Ethical Regulator Theorem to crises", *Acta Europæana Systemica*, 8: 53-58.
- Attenberg, J., P. Ipeirotis and F. Provost (2015). "Beat the machine: Challenging humans to find a predictive model's "unknown unknowns", *Journal of Data and Information Quality (JDIQ)*, 6(1): 1-17.
- Babcock, J., J. Kramár and R. Yampolskiy (2016). "The AGI containment problem", *International Conference on Artificial General Intelligence*, Cham: Springer.
- Babcock, J., J. Kramar and R. V. Yampolskiy (2019). "Guidelines for Artificial Intelligence Containment" In Ali. E. Abbas (ed.) *Next-Generation Ethics: Engineering a Better Society*, Padstow: Cambridge University Press: 90-112.
- Baclawski, K. (2018). "The observer effect", *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, IEEE.
- Baggili, I. and V. Behzadan (2019). "Founding the domain of AI forensics", *arXiv preprint arXiv:1912.06497*.
- Bandy, J. (2021). "Problematic machine behavior: A systematic literature review of algorithm audits", *Proceedings of the acm on human-computer interaction*, 5 (CSCW1): 1-34.
- Belrose, N., Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman and J. Steinhardt (2023). "Eliciting Latent Predictions from Transformers with the Tuned Lens", *arXiv preprint arXiv:2303.08112*.
- Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut and E. Brunskill (2021). "On the opportunities and risks of foundation models", *arXiv preprint arXiv:2108.07258*.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*, New York: Oxford University Press.

- Bowman, S. R., J. Hyun, E. Perez, E. Chen, C. Pettit, S. Heiner, K. Lukosuite, A. Askeell, A. Jones and A. Chen (2022). "Measuring progress on scalable oversight for large language models", *arXiv preprint arXiv:2211.03540*.
- Brcic, M. and R. V. Yampolskiy (2023). "Impossibility Results in AI: a survey", *ACM Computing Surveys*.
- Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li and S. Lundberg (2023). "Sparks of artificial general intelligence: Early experiments with gpt-4", *arXiv preprint arXiv:2303.12712*.
- Burnell, R., W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein and M. Mitchell (2023). "Rethink reporting of evaluation results in AI", *Science*, 380(6641): 136-138.
- Burns, C., H. Ye, D. Klein and J. Steinhardt (2022). "Discovering latent knowledge in language models without supervision", *arXiv preprint arXiv:2212.03827*.
- Busch, P., T. Heinonen and P. Lahti (2007). "Heisenberg's uncertainty principle", *Physics reports*, 452(6): 155-176.
- Caballero, E., K. Gupta, I. Rish and D. Krueger (2022). "Broken Neural Scaling Laws", *arXiv preprint arXiv:2210.14891*.
- Chalmers, D. J. (2023). "Could a large language model be conscious?", *arXiv preprint arXiv:2303.07103*.
- Chandrasekaran, V., H. Jia, A. Thudi, A. Travers, M. Yaghini and N. Papernot (2021). "SoK: Machine learning governance", *arXiv preprint arXiv:2109.10870*.
- Chen, F. and G. Roşu (2003). "Towards monitoring-oriented programming: A paradigm combining specification and implementation", *Electronic Notes in Theoretical Computer Science*, 89(2): 108-127.

- Christiano, P. (June 3, 2020). *Inaccessible Information*. Available at: <https://www.alignmentforum.org/posts/ZyWyAJbedvEgRT2uF/inaccessible-information>.
- Christiano, P., A. Cotra and M. Xu (2021). *Eliciting latent knowledge: How to tell if your eyes deceive you*.
- Clark, A. and D. Chalmers (1998). "The extended mind", *Analysis*, 58(1): 7-19.
- Council, N. R. (2005). *Monitoring Nuclear Weapons and Nuclear-Explosive Materials: An Assessment of Methods and Capabilities*, National Academies Press.
- Critch, A. and D. Krueger (2020). "AI research considerations for human existential safety (ARCHES)", *arXiv preprint arXiv:2006.04948*.
- Cummings, M. L. (2017). "Automation bias in intelligent time critical decision support systems", *Decision making in aviation*, New York: Routledge, pp. 289-294.
- Daras, G. and A. G. Dimakis (2022). "Discovering the hidden vocabulary of dalle-2", *arXiv preprint arXiv:2206.00169*.
- Elamrani, A. and R. V. Yampolskiy (2019). "Reviewing tests for machine consciousness", *Journal of Consciousness Studies*, 26(5-6): 35-64.
- Falco, G., B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta and C. Hart (2021). "Governing AI safety through independent audits", *Nature Machine Intelligence*, 3(7): 566-571.
- Gabriel, I. (2020). "Artificial intelligence, values, and alignment", *Minds and machines*, 30(3): 411-437.
- Ganguli, D., D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain and N. Elhage (2022). "Predictability and surprise in large generative models", *2022 ACM Conference on Fairness, Accountability, and Transparency*.

- Garrabrant, S., T. Benson-Tilsen, A. Critch, N. Soares and J. Taylor (2016). "Logical induction", *arXiv preprint arXiv:1609.03543*.
- Gibson, J. (1977). "The Theory of Affordances" In R. E. Shaw and J. Bransford (eds.), *Perceiving, Acting, and Knowing*, Hillsdale: Lawrence Erlbaum Associates.
- Goertzel, B. (2012). "Should humanity build a global AI nanny to delay the singularity until it's better understood?", *Journal of consciousness studies*, 19(1-2): 96-111.
- Goldwasser, S., M. P. Kim, V. Vaikuntanathan and O. Zamir (2022). "Planting undetectable backdoors in machine learning models", *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS), IEEE*.
- Gutierrez, C. I. (2020). "The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the US."
- Hendrycks, D., N. Carlini, J. Schulman and J. Steinhardt (2021). "Unsolved problems in ml safety", *arXiv preprint arXiv:2109.13916*.
- Hendrycks, D. and M. Mazeika (2022). "X-risk analysis for ai research", *arXiv preprint arXiv:2206.05862*.
- Hermann, R. and A. Krener (1977). "Nonlinear controllability and observability", *IEEE Transactions on automatic control*, 22(5): 728-740.
- Holmström, B. (1979). "Moral hazard and observability", *The Bell journal of economics*: 74-91.
- Honovich, O., U. Shoham, S. R. Bowman and O. Levy (2022). "Instruction induction: From few examples to natural language task descriptions", *arXiv preprint arXiv:2205.10782*.
- Horowitz, M. C. (2019). "When speed kills: Lethal autonomous weapon systems, deterrence and stability", *Journal of Strategic Studies*, 42(6): 764-788.

- Horvitz, E. (2016). One hundred year study on artificial intelligence, Stanford University.
- Howe, W. and R. Yampolskiy (2021). "Impossibility of Unambiguous Communication as a Source of Failure in AI Systems", *AISafety@IJCAI*.
- Johnson, N., G. Zhao, E. Hunsader, J. Meng, A. Ravindar, S. Carran and B. Tivnan (2012). "Financial black swans driven by ultrafast machine ecology", *arXiv preprint arXiv:1202.1448*.
- Johnson, N., G. Zhao, E. Hunsader, H. Qi, N. Johnson, J. Meng and B. Tivnan (2013). "Abrupt rise of new machine ecology beyond human response time", *Scientific Reports*, 3(1): 2627.
- Kambhampati, S. (2022). "Changing the nature of AI research", *Communications of the ACM*, 65(9): 8-9.
- Krener, A. J. and K. Ide (2009). "Measures of unobservability", *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, IEEE*.
- Kurakin, A., I. Goodfellow and S. Bengio (2016). "Adversarial examples in the physical world", *arXiv preprint arXiv:1607.02533*.
- Levin, J.-C. and M. M. Maas (2020). "Roadmap to a Roadmap: How Could We Tell When AGI is a 'Manhattan Project' Away?", *arXiv preprint arXiv:2008.04701*.
- Lipton, R. J., K. W. Regan, R. J. Lipton and K. W. Regan (2013). "David Johnson: Galactic algorithms." *People, Problems, and Proofs: Essays from Gödel's Lost Letter: 2010*: 109-112.
- Liu, Y.-Y., J.-J. Slotine and A.-L. Barabási (2013). "Observability of complex systems", *Proceedings of the National Academy of Sciences*, 110(7): 2460-2465.
- Logan, D. C. (2009). "Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry", *Journal of experimental botany*, 60(3): 712-714.

- Lyon, D. (2007). "Surveillance studies: An overview."
- Mialon, G., R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu and A. Celikyilmaz (2023). "Augmented language models: a survey", *arXiv preprint arXiv:2302.07842*.
- Michaud, E. J., Z. Liu, U. Girit and M. Tegmark (2023). "The quantization model of neural scaling", *arXiv preprint arXiv:2303.13506*.
- Mökander, J. and L. Floridi (2021). "Ethics-based auditing to develop trustworthy AI", *Minds and Machines*, 31(2): 323-327.
- Newman, J. (January 2023). "A Taxonomy of Trustworthiness for Artificial Intelligence", *CLTC White Paper Series*.
- Nguyen, D., S. Yingchareonthawornchai, V. Tekken Valapil, S. S. Kulkarni and M. Demirbas (2021). "Precision, recall, and sensitivity of monitoring partially synchronous distributed programs", *Distributed Computing*, 34: 319-348.
- OpenAI (2023). "GPT-4 System Card" Available at: <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.
- OpenAI (2023). "GPT-4 technical report." *arXiv*.
- Ornes, S. (March 16, 2023). "The Unpredictable Abilities Emerging From Large AI Models", *Quanta Magazine*. Available at: <https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316>.
- Ortega, P. A. and V. Maini (September 27, 2018). "Building safe artificial intelligence: specification, robustness, and assurance". Available at: <https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1>.
- Pedersen, T. and C. Johansen (2020). "Behavioural artificial intelligence: an agenda for systematic empirical studies of artificial inference", *AI & Society*, 35(3): 519-532.

- Perez, E., S. Ringer, K. Lukošiušė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu and S. Kadavath (2022). "Discovering Language Model Behaviors with Model-Written Evaluations", *arXiv preprint arXiv:2212.09251*.
- Power, A., Y. Burda, H. Edwards, I. Babuschkin and V. Misra (2022). "Grokking: Generalization beyond overfitting on small algorithmic datasets", *arXiv preprint arXiv:2201.02177*.
- Rahwan, I., M. Cebrian, N. Obradovich, J. Bongard, J. F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin and M. O. Jackson (2022). "Machine behaviour", *Machine Learning and the City: Applications in Architecture and Urban Design*: 143-166.
- Raji, I. D., A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron and P. Barnes (2020). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing", *Proceedings of the 2020 conference on fairness, accountability, and transparency*.
- Ramamoorthy, A. and R. Yampolskiy (2017). "Beyond Mad?: The Race for Artificial General Intelligence", *ITU Journal: ICT Discoveries*.
- Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature machine intelligence*, 1(5): 206-215.
- Sandberg, A. (2019). "There is plenty of time at the bottom: The economics, risk and ethics of time compression", *Foresight*, 21(1): 84-99.
- Saunders, W., G. Sastry, A. Stuhlmüller and O. Evans (2017). "Trial without error: Towards safe reinforcement learning via human intervention", *arXiv preprint arXiv:1707.05173*.
- Schaeffer, R., B. Miranda and S. Koyejo (2023). "Are Emergent Abilities of Large Language Models a Mirage?", *arXiv preprint arXiv:2304.15004*.

- Schneider, J. and F. Breiting (2020). "AI Forensics: Did the artificial intelligence system do it? why?", *arXiv preprint arXiv:2005.13635*.
- Schwarting, M., T. Burton and R. Yampolskiy (2015). "On the Obfuscation of Image Sensor Fingerprints", *Information and Computer Technology (GOCICT), 2015 Annual Global Online Conference on, IEEE*.
- Scott, P. J. and R. V. Yampolskiy (2019). "Classification schemas for artificial intelligence failures", *Delphi*, 2: 186.
- Shavit, Y. (2023). "What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring", *arXiv preprint arXiv:2303.11341*.
- Shevlane, T., S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano and A. Dafoe (May 24, 2023). "Model evaluation for extreme risks". Available at: <https://arxiv.org/pdf/2305.15324.pdf>.
- Sotala, K. (2012). "Advantages of artificial intelligences, uploads, and digital minds", *International journal of machine consciousness*, 4(01): 275-291.
- Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shueb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta and A. Garriga-Alonso (2022). "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models", *arXiv preprint arXiv:2206.04615*.
- Tabassi, E. (2023). "Artificial Intelligence Risk Management Framework (AI RMF 1.0)."
- Tran, B., J. Li and A. Madry (2018). "Spectral signatures in backdoor attacks", *Advances in neural information processing systems*, 31.
- Trazzi, M. and R. V. Yampolskiy (2020). "Artificial Stupidity: Data We Need to Make Machines Our Equals," *Patterns*, 1(2): 100021.

- Turchin, A. (2021). "Catching Treacherous Turn: A Model of the Multilevel AI Boxing". Available at: https://www.researchgate.net/profile/Alexey-Turchin/publication/352569372_Catching_Treacherous_Turn_A_Model_of_the_Multilevel_AI_Boxing.
- van Leeuwen, J. and J. Wiedermann (2021). "Impossibility results for the online verification of ethical and legal behaviour of robots", *Utrecht University, Utrecht, UU-PCS-2021-02*.
- Vance, A. (May 15, 2022). "Is AI Progress Impossible To Predict?". Available at: <https://www.lesswrong.com/posts/G993PFTwqqdQv4eTg/is-ai-progress-impossible-to-predict>.
- Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou and D. Metzler (2022). "Emergent abilities of large language models", *arXiv preprint arXiv:2206.07682*.
- Wei, J., X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le and D. Zhou (2022). "Chain of thought prompting elicits reasoning in large language models", *arXiv preprint arXiv:2201.11903*.
- Werkhoven, P., L. Kester and M. Neerincx (2018). "Telling autonomous systems what to do", *Proceedings of the 36th European Conference on Cognitive Ergonomics*.
- Whittlestone, J. and J. Clark (2021). "Why and How Governments Should Monitor AI Development", *arXiv preprint arXiv:2108.12427*.
- Wiener, N. (1960). "Some moral and technical consequences of automation", *Science*, 131(3410): 1355-1358.
- Wolfram, S. (2002). *A new kind of science*, Champaign: Wolfram Media.
- Woolgar, S. (1985). "Why not a sociology of machines? The case of sociology and artificial intelligence", *Sociology*, 19(4): 557-572.
- Wu, F. F. and A. Monticelli (1985). "Network observability: theory", *IEEE Transactions on Power Apparatus and Systems*, (5): 1042-1048.

- Yampolskiy, R. (2018). "The Singularity May Be Near", *Information*, 9(8): 190.
- Yampolskiy, R. (2020). "On controllability of artificial intelligence", *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*.
- Yampolskiy, R. (2021). "On the Differences between Human and Machine Intelligence", *AISafety@IJCAI*.
- Yampolskiy, R. V. (2008). "Behavioral modeling: an overview", *American Journal of Applied Sciences*, 5(5): 496-503.
- Yampolskiy, R. V. (2012). "Leakproofing Singularity - Artificial Intelligence Confinement Problem", *Journal of Consciousness Studies (JCS)*, 19(1-2): 194-214.
- Yampolskiy, R. V. (2015). "Analysis of types of self-improving software", *Artificial General Intelligence: 8th International Conference, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings 8, Springer*.
- Yampolskiy, R. V. (2017). "What are the ultimate limits to computational techniques: verifier theory and unverifiability", *Physica Scripta*, 92(9): 093001.
- Yampolskiy, R. V. (2018). "Artificial consciousness: An illusionary solution to the hard problem", *Reti, saperi, linguaggi*, (2): 287-318.
- Yampolskiy, R. V. (2020). "Unexplainability and Incomprehensibility of AI", *Journal of Artificial Intelligence and Consciousness*, 7(2): 277-291.
- Yampolskiy, R. V. (2020). "Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent", *Journal of Artificial Intelligence and Consciousness*, 7(1): 109-118.
- Yampolskiy, R. V. (2022). "AI Risk Skepticism", *Philosophy and Theory of Artificial Intelligence 2021, Springer*: 225-248.
- Yampolskiy, R. V. (2022). "Metaverse: A Solution to the Multi-Agent Value Alignment Problem", *Journal of Artificial Intelligence and Consciousness*: 1-11.

- Yampolskiy, R. V. (2022). "On the Controllability of Artificial Intelligence: An Analysis of Limitations", *Journal of Cyber Security and Mobility*: 321–404–321–404.
- Yampolskiy, R. V. (2023). "Ownability of AGI", *Artificial General Intelligence: 15th International Conference, AGI 2022, Seattle, WA, USA, August 19–22, 2022, Proceedings*, Springer.
- Yampolskiy, R. V. (2024). *AI - Unpredictable, Unexplainable, Uncontrollable*, Florida: CRC Press.
- Yampolskiy, R. V., L. Ashby and L. Hassan (2012). "Wisdom of artificial crowds—a metaheuristic algorithm for optimization."
- Yampolskiy, R. V. and V. Govindaraju (2006). "Use of behavioral biometrics in intrusion detection and online gaming", *Biometric Technology for Human Identification III, SPIE*.
- Yudkowsky, E. (May 22, 2008). "That Alien Message", *Less Wrong*. Available at: <https://www.lesswrong.com/posts/5wMcKNAwB6X4mp9og/that-alien-message>.
- Yudkowsky, E. (October 13, 2017). "There's No Fire Alarm for Artificial General Intelligence". Available at: <https://intelligence.org/2017/10/13/fire-alarm/>.
- Yudkowsky, E. and R. Hanson (2008). "The Hanson-Yudkowsky AI-foom debate", *MIRI Technical Report*. Available at: <http://intelligence.org/files/AIFoomDebate.pdf>.
- Zhang, D., S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J. C. Niebles and M. Sellitto (2021). "The AI index 2021 annual report", *arXiv preprint arXiv:2103.06312*.
- Zhou, D., N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le and E. Chi (2022). "Least-to-most prompting enables complex reasoning in large language models", *arXiv preprint arXiv:2205.10625*.

Ziesche, S. and R. V. Yampolskiy (2020). "Towards the Mathematics of Intelligence", *The Age of Artificial Intelligence: An Exploration*, 1.

Zoe Cremer, C. and J. Whittlestone (2021). "Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI".

Short Bios

Ana Rita Pina holds a Master's in Portuguese as a Second Language/ Foreign Language from the University of Porto as well as a Bachelor's in Translation. As of the publishing of this paper, she works as a PFL teacher.

Antonino Drago is a retired Associate Professor from the University "Federico II" of Naples, Italy, where he served in the Department of Physics. He earned his Laurea in Physics from the University of Pisa in 1961 and subsequently held research fellowships in biophysics and cybernetics in both Pisa and Naples. His academic career spans several decades, during which he taught courses in physics, epistemology, and the history of science. Drago is the author of six monographs, editor of nine scholarly volumes, and has published over 400 scientific articles along with 200 additional contributions in peace research and applied ethics.

Cosimo Palma is a PhD Candidate at the University of Naples "L'Orientale" and the University of Pisa, within the Italian national doctoral program in Artificial Intelligence. His academic background ranges over Philosophy, History, Computational Linguistics and Informatics, whereas his main research project deals with the integration of Large Language Models and Knowledge Graphs for Automatic Story Generation in the domain of Cultural Heritage. During his research visit at the C2DH department of the University of Luxembourg, he has worked on data and tools harmonization for Digital History workflows.

Daniela Alexandra Rodrigues holds a Master's in Translation by the University of Lisbon. Since then, she has participated in Machine

Translation and Machine Translation Literacy research projects at CETAPS and has been researching translation and translation technologies.

Denis Coitinho Is Full Professor of Philosophy at University of Vale do Rio dos Sinos (Brazil) and visiting scholar at University of Oxford (UK) and Harvard University (USA). PhD in Philosophy from the Pontifical Catholic University of Rio Grande do Sul (PUCRS). He is a Brazilian National Research Council (CNPq) and author of *Contrato & Virtudes* I (Loyola, 2016) e II (Loyola, 2021) and of the chapter “The COVID-19 pandemic and the language of virtues” in the book *Moral Challenges in a Pandemic Age*, edited by Evandro Barbosa (Routledge, 2023).

Eleni Angelou is a PhD student at CUNY, Graduate Center. Her research focuses on the philosophy of science, especially metaphysics and theory of science, and cognitive science. She is also interested in the history of philosophy (mostly David Hume and American transcendentalists and pragmatists). She is the translator of Thoreau’s Letters to Harrison Blake in Greek.

Germán Massaguer Gómez is an PhD researcher and philosopher specializing in human-robot interaction and AI ethics, currently affiliated with Universidad Carlos III de Madrid. His work critically examines the ethical implications of trust and moral consideration in relations between humans and social robots.

Koray Karaca is an assistant professor in the Department of Philosophy at the University of Twente in the Netherlands. He holds a Ph.D. in theoretical physics from the Middle East Technical University, Ankara, Turkey, and a Ph.D. in history and philosophy of science from Indiana University, Bloomington, USA. Earlier, he worked at the University of Wuppertal, Germany, and the University

of South Florida, USA. His research interests include the epistemology of scientific experimentation, scientific modelling, and machine learning.

Luiza Cunha graduated in Law and has a PhD in Law and Political Science at the University of Barcelona, line of research in Bioethics and Law. Has a master's degree at the University of Lisbon in Law and Juridical Science, line of research in International Juridical Science. Public Federal employee at the Ministry of Science, Technology and Innovation from Brazil.

Marcela Castanheira graduated and has a master's degree in Philosophy at the Federal University of Goiás. Has a PhD in Philosophy at the Federal University of Minas Gerais. Philosophy teacher at Goiás' Federal Institute of Education, Science and Technology. Researcher with experience on Ethics and Political Philosophy, with an emphasis on Foucault's work.

Marta Mendonça is an Associate Professor in the Philosophy Department at Universidade Nova de Lisboa, where she coordinates the Philosophy degree programme. She has a degree in Philosophy from the Universidad de Navarra and a Master's and PhD from the Universidade Nova de Lisboa (2001). She studied in Spain (Navarra), France (Paris 1, Paris 4 and EHESS), the United Kingdom (Oxford) and Germany (Berlin). She publishes mainly in the fields of Modern Philosophy, Philosophy of Science, Philosophy of Nature and Bioethics. She has worked as a visiting professor at various universities in Spain, France, Brazil, Chile and the UK. She is a researcher at Ifilnova at NOVA FCSH.

Pablo Muñoz is a PhD Candidate in the Department of Philosophy at University of the Andes in Colombia and works as Innovation Management Professional at Compensar. He holds a MSc degree in

Philosophy of Science, Technology and Society from the University of Twente and has worked in innovation and project management in the banking, wellness and healthcare industries. His research interests include the epistemology of machine learning, scientific modelling, and science and human values.

Parashar Das is a PhD Candidate in Public International Law at LSE Law School. He completed his undergraduate legal training at University of Adelaide before pursuing theoretical questions across human rights, judicial reasoning, medical ethics and international law & international relations in his LLM program at University of Melbourne. Prior to LSE Law, Parashar pursued journalism in international affairs, Aboriginal issues and asylum seeker detention practices.

Roman V. Yampolskiy is an Associate Professor of Computer Engineering & Computer Science at the Speed School of Engineering, University of Louisville, where he directs the Cyber Security Lab. He earned his PhD in Computer Science & Engineering from the University at Buffalo in 2008 and is senior member of IEEE and AGI. Dr. Yampolskiy is a pioneer in AI safety and cybersecurity, best known for his research on AI predictability, explainability, and controllability. He is author/editor of over a hundred publications and several books – including *Artificial Superintelligence: A Futuristic Approach* and *Artificial Intelligence Safety and Security* – and has received numerous awards for teaching excellence and research innovation.

Steven S. Gouveia earned his PhD in Neurophilosophy of Mind from the University of Minho under the supervision of Manuel Curado and Georg Northoff. He currently serves as a Research Fellow leading a six-year research project on the Ethics of AI in Medicine at the Mind, Language and Action Group of the Institute of Philosophy of the University of Port. He also held an Honorary Professorship at Andrés

Bello University's Faculty of Medicine in Chile since June 2023. He is the author and editor of 14 peer-reviewed academic books, including *Philosophy & Neuroscience: a Methodological Analysis* (2022), *Predictive Minds: Old Problems and New Challenges* (2023), and *The Odyssey of the Mind: Dialogues on the Brain and Consciousness* (2024), which features conversations with thinkers like David Chalmers, Sir Roger Penrose, Anil Seth and others.

Yinchun Wang is a Postdoctoral Fellow in Philosophy, Associate Professor and Master's Tutor of Donghua University, Visiting Scholar at University College London (UCL) under the auspices of the China Scholarship Council in 2016.