

FACULTY OF ENGINEERING OF UNIVERSITY OF PORTO

Artificial Intelligence and Infodemic: a study on fact-checked Health Communication and Synthetic Media

Haline Maia



Doctoral Program in Digital Media

Supervisors:

Helena Lima, Faculty of Arts and Humanities of the University of Porto
Sérgio Nunes, Faculty of Engineering of the University of Porto

February 24, 2025

**Artificial Intelligence and Infodemic:
a study on fact-checked Health Communication
and synthetic media**

Haline Maia

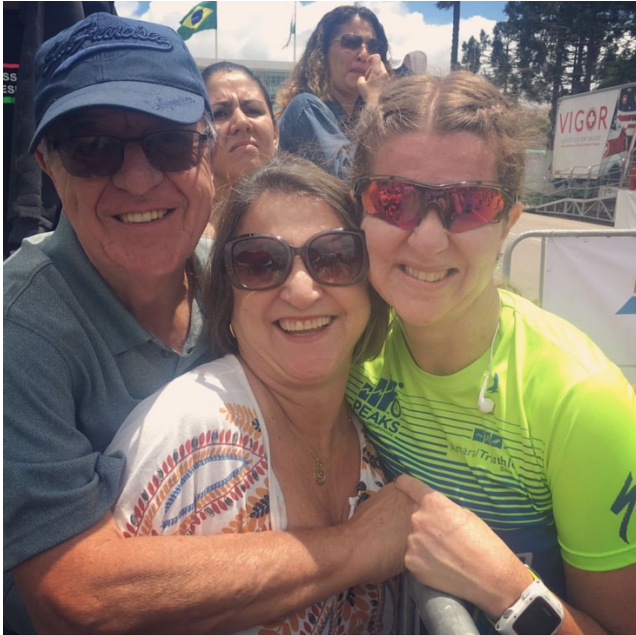
Doctoral Program in Digital Media

February 24, 2025



This study was financed by national funds through the FCT - Foundation for Science and Technology, through the award of a doctoral scholarship (2021.05707.BD).

Aos meus pais, Soely Costa Maia e Carlos Sthalshmidt Maia,
sempre torcendo por mim em todas as maratonas da vida.¹



© Finish line at Curitiba Marathon, November 2017.

¹ Translation to English: To my parents, Soely Costa Maia and Carlos Sthalshmidt Maia, always cheering for me in all the marathons of life.

ABSTRACT

The proliferation of health misinformation, especially during critical times such as pandemics, has underscored the need for effective mechanisms to verify and disseminate accurate health news. This research, guided by Value Sensitive Design (VSD), investigates the integration of Artificial Intelligence (AI) in health fact-checking, aiming to enhance the speed and reliability of information dissemination while ensuring ethical compliance.

The study addresses these key research questions: How can AI facilitate the rapid dissemination of authenticated health information? What benefits does AI integration bring to health fact-checking processes? How can AI promote ethically responsible practices in the dissemination of counter-information? Through systematic reviews, case studies, and empirical research, including co-design workshops and surveys, the research evaluates existing AI applications. It develops guidelines for incorporating AI in journalism and public health information systems.

This dissertation follows a structured approach centered around the three distinct phases of VSD. During the Conceptual Investigations phase, systematic reviews and meta-analyses were conducted on publications from 2020 to 2022, using the PRISMA process to analyze 57 studies. During the Technical Investigations phase, case studies and semi-structured interviews were conducted with stakeholders. In the Empirical Investigations, technological probes using AI for the dissemination phase of health fact-checking were tested through co-design workshops, focus groups, and surveys. The data collected in the Technical and Empirical phases were analyzed using thematic and exploratory methods.

Findings indicate that while AI significantly enhances the efficiency of fact-checking processes, challenges related to equality, governance, and stakeholder trust remain prevalent. The study also explores the socio-technical dynamics of AI applications in fact-checking, emphasizing the importance of value-driven design to mitigate ethical risks and promote inclusivity.

The implications of this research are far-reaching, offering guidelines for developing AI-driven tools that are not only technologically effective but also

culturally sensitive and ethically sound. By fostering a better understanding of AI's role in managing health misinformation, this work contributes to the broader discussion on technology governance and the ethical dimensions of digital media in public health contexts.

Keywords: AI Ethics; Infodemic; Generative Media; Fact-Checking; Health Misinformation; Media Literacy; Journalism Innovation

RESUMO

A proliferação de desinformação sobre saúde, especialmente em tempos críticos como pandemias, destacou a necessidade de mecanismos eficazes para verificar e disseminar notícias de saúde precisas. Esta pesquisa, orientada pela metodologia *Value Sensitive Design* (VSD), investiga a integração da Inteligência Artificial (IA) no processo de verificação de fatos em saúde, visando melhorar a velocidade e a confiabilidade da disseminação de informações, garantindo ao mesmo tempo a conformidade ética.

O estudo aborda as seguintes questões-chave de pesquisa: Como a IA pode facilitar a rápida disseminação de informações de saúde autenticadas? Quais benefícios a integração da IA traz aos processos de verificação de fatos em saúde? Como a IA pode promover práticas eticamente responsáveis na disseminação de contrainformação? Por meio de revisões sistemáticas, estudos de caso e pesquisas empíricas, incluindo workshops de co-design e levantamentos, a pesquisa avalia as aplicações existentes de IA e desenvolve diretrizes para incorporar a IA no jornalismo e nos sistemas de informação de saúde pública.

Esta dissertação adota uma abordagem estruturada em três fases do VSD. Na fase de Investigações Conceptuais, foram realizadas revisões sistemáticas e meta-análises de publicações entre 2020 e 2022, utilizando o processo PRISMA para analisar 57 estudos. Nas Investigações Técnicas, foram conduzidos estudos de caso e entrevistas semi-estruturadas com cidadãos. Nas Investigações Empíricas, foram testadas experiências tecnológicas com uso de IA para a fase de divulgação de checagem de factos de saúde, através de workshops de co-design, grupos focais e inquéritos. Os dados recolhidos nas fases Técnicas e Empíricas foram analisados através de análise temática e abordagens exploratórias.

Os resultados indicam que, embora a IA melhore significativamente a eficiência dos processos de verificação de fatos, desafios relacionados à igualdade, governança e confiança dos entrevistados permanecem prevalentes. O estudo também explora as dinâmicas socio-técnicas das aplicações de IA em checagem

de factos, enfatizando a importância do design orientado por valores para mitigar riscos éticos e promover a inclusividade.

As implicações desta pesquisa são abrangentes, oferecendo diretrizes para o desenvolvimento de ferramentas impulsionadas por IA que sejam não apenas tecnologicamente eficazes, mas também culturalmente sensíveis e eticamente sólidas. Ao fomentar uma melhor compreensão do papel da IA na gestão da desinformação em saúde, este trabalho contribui para a discussão mais ampla sobre governança tecnológica e as dimensões éticas da mídia digital em contextos de saúde pública.

Keywords: Ética da IA; Infodemia; Mídia Generativa; Verificação de Fatos; Desinformação em Saúde; Letramento Midiático; Inovação no Jornalismo

ACKNOWLEDGMENTS

Acknowledgements represent the most gratifying yet emotionally challenging moment of this dissertation, which was born from a dream of moving to another country to pursue a doctorate, threatened during pandemic times but completed in four very intense years, transitioning from social distancing to the return to “normal” life. Countless individuals have traversed this journey with me, from its conception to its completion. There are those who remained steadfast companions throughout the entire marathon, while others, equally significant, appeared at various specific moments. Some contributed to portions and parts of the research, others engaged in vibrant exchanges of ideas, and there are those who helped uplift my spirit during moments of discouragement and sadness. To each one who extended a helping hand in their unique way, I extend my heartfelt gratitude. I hope to pay these acts of kindness forward.

The dream of this thesis would have remained a nightmare if it were not for the support of the Foundation for Science and Technology (FCT). The funding of the research project and the acceptance for the completion of two mobilities, a short one to Utrecht University and another scientific expedition of four months to Lund University, made this goal possible. The Portuguese Academy stood by me during the adaptation phases to their country and a global pandemic, showing acceptance and valuing my research project.

Over the years, I have been fortunate to have exceptional mentors who believed in the potential of this thesis, even when it existed only as a brief plan awaiting development. Special thanks go, as always, to Helena Lima, whose constant friendship and acceptance of my ambitions have never wavered. I also thank my co-supervisor, Sérgio Nunes, for helping me with his objectivity to find anchorage and scientific relevance. Christopher Mathieu, my mentor during my time at Lund University, offered his friendship and unwavering support in every possible way. I am equally grateful to my remarkable colleagues at the Faculty of Engineering of the University of Porto and all those I have crossed paths within the Doctoral Programme in Digital Media (PDMD).

Within the scope of the PDMD, I would like to give special thanks to Marisa Silva, responsible for the secretariat, who makes a difference with her collaboration. Furthermore, I extend my sincere gratitude to the Director of PDMD, António Coelho, whose fundamental and constant support propelled several activities carried out by me at the Faculty of Engineering of the University of Porto. Jordana Casarin, my southern compatriot, became more than my colleague in PDMD, always present in necessary trips and a collector of unforgettable moments, a friendship I intend to carry for life. I also thank my colleague at Infolab, Inês Koch, always up for a good conversation and willing to help me with the bureaucracies of the academic journey. From Infolab, I move on to thank the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) and its fellows for providing a supportive environment.

The PDMD also brought me another very special friendship with Patrícia Cordeiro, who welcomed me into her home in Macedo and made me feel at home

there too. Through Patrícia's kindness also came the opportunity to be part of my research with the EducAge project, to whom I extend thanks to all who authorised the completion of this joint work, as well as those who allowed themselves to contribute and give their valuable opinions, which are reported in this thesis. I also thank Priscila Ariel, my collaborator and great inspiration as a researcher and technology developer. From her, I also highlight my thanks to Globo, a partnership that started during my master's and continued in my PhD, due to my previous professional history achieved in their spaces.

My gratitude also extends to my friends inside and outside academia, who consistently provided a support base for both work-related matters and life in general. I start by thanking my friends in Brazil, who, even from afar, always made themselves present. Gisele Obara, Ricardo Bois, Denise Chybior, Anaelse Oliveira, Simone Dias, Lucrecia Aida Carvalho, Roniane Correia de Freitas, Rogério Galindo, Ana Johann, Renato Essenfelder and Catarina, Karen Czachorowski thank you for being present, even at a distance, during the pandemic years, the first of my PhD, in which I often thought of giving up. Your presence, even if virtual, was fundamental for maintaining my mental health during this journey. Thank you also for always visiting me in Brazil; some even came to see me in Portugal, like my dear sister Gi. I hope the content of this research makes up for the pain of not having you always close by. And also thank my dear friend Priscila Pinto, from Curitiba to Rio, to Madrid and everywhere I go, she is always present and cheering for me.

Thank you to all the friends I made in the beautiful city of Porto who accompanied me during each stage of this project: to my admirable confidant Priscilla Brasil and her lovely family. To the most beloved couple, Alexandre and Jacinto, who, each in their own way, were always a comfort. To my friend Natália Córdova. I also extend my immense gratitude to the dear friends I made during my short time in Lund. My thanks go to my friends Stuti, Gautam, and those who became my family in Lund, introduced to me by my host Chris, Ossian, Gonçalo, and Nicole. Not forgetting the coffee machine in the sociology department, a great companion while finalizing the writing of the thesis.

Finally, my debt of gratitude to my family is immeasurable, as they have always been with me through all the difficulties I have faced so far. My parents, Soely and Carlos, dedicated themselves tirelessly to help make my dream of a doctorate a reality. In my academic marathon, as in the others I have done before, I know they were anguished at times with their youngest daughter on the other side of the Atlantic, facing lion-hearted evaluators and the SARS-CoV virus. Nevertheless, they always encouraged me to seek the best and made themselves available to listen to me in video calls and even to visit me when possible. To my siblings Júnior and Karina, as well as my in-laws Cassiana and Rubens, thank you for your affection and friendship. To my dear nephews, your aunt loves you very much and will always miss you. As always, thank you Larissa, my cousin, my soulmate. I am eternally grateful for your affection and infinite friendship. To all my other uncles, aunts, cousins, and dear relatives who always welcome me with joy at the end of the year, my heartfelt thanks as well.

Finally, thank you to all the internet users, misinformed or well-informed, memers, tech-users, techno skeptics, techno-illiterates, and life beings who have shaped the digital sphere.



O Mez da Grippe,
Valêncio Xavier (1933 – 2008)

² Illustration from a literary work originally published in 1981 tells the story of the impact of the Spanish flu on Curitiba, Brazil, in 1918. The work uses newspaper cuttings, adverts, engravings, testimonies, and fictional interventions. The following is a free translation of the above: Beware of the Spanish flu! Use Santa Helena Balm, a powerful antiputrid, analgesic disinfectant, and enemy of foul odors! Used for gargling against bad breath and sore throats. Available for 1\$500 in every pharmacy. Only St Helena's Balm.

CONTENTS

CHAPTER 1: INTRODUCTION	11
1. PROBLEM AND MOTIVATION.....	12
2. OBJECTIVES OF THE RESEARCH.....	14
3. HYPOTHESIS AND RESEARCH QUESTIONS.....	15
4. NARROWING THE TOPIC: TECHNOLOGY APPLIED TO HEALTH FACT-CHECKING	17
5. OVERVIEW OF THE THESIS STRUCTURE.....	18
CHAPTER 2: REVIEW OF THE STATE OF THE ART	21
SECTION 1: FROM DIGITAL ECOSYSTEM TO HEALTH DISINFORMATION AND FACT-CHECKING. 21	
1.1 Digital ecosystem and technology	21
1.2 Post-truth and mediation.....	29
1.3 Disinfodemic perspective and challenges	36
1.4 Fact-checking as a way for accountability in journalism	41
SECTION 2: SYNTHETIC MEDIA CONTRIBUTION TO SHARE HEALTH FACT-CHECKING	50
2.1 Artificial Intelligence Ethical Issues	51
2.2 Automation, Generative Media, and AI.....	73
2.3 Automatic Fact-checking	91
CHAPTER SUMMARY	106
CHAPTER 3: RESEARCH METHODOLOGY	107
SECTION 3. METHODOLOGICAL APPROACHES IN A VALUE SENSITIVE DESIGN STUDY	108
3.1 Phase 1 – Conceptual Investigations	111
3.2 Phase 2 – Technical Investigations	113
3.3 Phase 3 - Empirical Investigations	115
3.4 The stakeholders involved	124
3.5 Methodological Challenges and Mitigation Strategies	129
OUTLINE OF RESEARCH METHODOLOGY	130
CHAPTER 4: TECHNICAL AND EMPIRICAL RESULTS	133
SECTION 4. CASE STUDIES AND INTERVIEWS	134
4.1 Publishing Fact-checking Results	135
4.2 Virtual Characters and Deep Learning	143
SECTION 5. DEVELOPMENT OF CO-DESIGN WORKSHOPS	151
5.2 Workshop Unravelling Value Assessment Scenarios	156
5.3 VSD Methods in a Multi-Cultural Workshop.....	161
5.4 Value-oriented Workshop in Elderly Health Information	175
CHAPTER SUMMARY	208
CHAPTER 5: VALUE-FRAMEWORKS AND DISCUSSIONS	209
SECTION 6. VALUES, NORMS AND TECHNICAL REQUIREMENTS.....	210
6.1 Integrating values into the design process	212
6.2 Ethical Guidelines and Principles.....	222
6.3 Mock-up proposal	233
6.4 Limitations of the Proposed Framework	239
CHAPTER SUMMARY	242
CHAPTER 6: CONCLUSIONS	243
REFERENCES	259
ANNEXURES AND APPENDICES	287

LIST OF FIGURES

Figure 1. Intervention ecosystem proposal inspired by UNESCO Disinfodemic study	14
Figure 2. Continuity model of science communication	34
Figure 3. A Snapshot of the International COVID-19 Fact-Checking Ecosystem...44	
Figure 4. Example from value to requirements (M.T. van Steenberg, Cathelijn 2022).....	60
Figure 5. Lip-syncing example from a famous internet Deepfake video	68
Figure 6. Nancy Pelosi manipulated to appear intoxicated	69
Figure 7. Reuters Virtual Reporter Prototype	70
Figure 8. AI Kazakh’s Atameken Business News Anchor.....	70
Figure 9. Lu for an advertising campaign. Image credit: Disclosure/Magalu	72
Figure 10. Exemplification of algorithm generalization (Geitgey, 2017)	80
Figure 11. Parameters for differentiating classification and regression algorithms (Geitgey, 2017)	81
Figure 12. Differentiating supervised and unsupervised learning (LeCun, 2016) ..	82
Figure 13. Illustration of a Deep Learning model in the layers (Goodfellow et al., 2016).....	84
Figure 14. Example of LDC tokenization, which in this case uses the abbreviation	86
Figure 15. Fragment of a hierarchical ontology (Bird, Klein, and Loper 2009).....	87
Figure 16. Four phases required to perform Fact-checking (Saquete et al. 2020) .	92
Figure 17. Prototypical fact-checking pipeline (Barrón-Cedeño et al. 2020).....	93
Figure 18. The first graph represents the spread of misinformation, and the other showcases the spread of non-misinformation content. The data showcases the average number of times the source tweets are retweeted or quoted within 4-hour intervals (Mu et al. 2023).....	94
Figure 19. PRISMA 2020 flow diagram for automatic fact-checking new systematic reviews.....	98
Figure 20. Top 10 conferences with the screened automatic fact-checking during the pandemic period.....	100
Figure 21. Less studied reported techniques according to the review	104
Figure 22. TreeMap with keywords fact-checking during the pandemic period .	104
Figure 23. Keyword co-occurrences clusters with automatic fact-checking during the pandemic	105
Figure 24. VSD tripartite approach.....	109
Figure 25. Four main groups of stakeholders to be included in the research	125
Figure 26. Fact-check publication utilizing a visual element to denote the claim as unfounded. Source: Observador.....	135
Figure 27. A campaign crafted to disseminate fact-checked information regarding Covid-19.....	136
Figure 28. Fact-check assistant interface	139
Figure 29. TITAN revision for a fact-checking flow chart.....	142
Figure 30. AIDA artistic process (Ariel, 2023). Source: Globo	146
Figure 31. Presentation tones: (a) formal and (b) informal (Ariel, 2023).....	147

Figure 32. Analysis of the input audio stream (Ariel, 2023).....	149
Figure 33. GaB presents sports News. Source: Globo.....	150
Figure 34. Storyboarding used in a participatory workshop with journalists to explore the evolution of fake news methods alongside AI translation tools.....	152
Figure 35. Basic Fact-checking Vídeo Creation FlowChart	154
Figure 36. Value Cards used in the workshop as a tool from VSD	158
Figure 37. Envisioning Card used in the workshop as a tool from VSD	163
Figure 38. Presentation of the initial concepts of the workshop.....	164
Figure 39. Developed thematic map, showing three main approaches and eight codes.....	170
Figure 40. Group #1 experience.....	172
Figure 41. Group #2 experience.....	173
Figure 42. Group #3 experience.....	174
Figure 43. Iterated flowchart for creating fact-checking videos	177
Figure 44. High-fidelity non-functional prototype with probes for EducAge site	180
Figure 45. Questionnaire layout with response scale symbols display.....	181
Figure 46. Grades assigned to feeling protected from each participant	184
Figure 47. Grades assigned to solutions from each participant in the 'Trust' category	185
Figure 48. Grades assigned to solutions from each participant in the 'Engagement' category	187
Figure 49. Grades assigned in the “engagement” category regarding resonance	188
Figure 50. Grades assigned to solutions from each participant about “Frustration”	189
Figure 51. Grades assigned to solutions from each participant about “Frightening”	190
Figure 52. Grades assigned to solutions from each participant in the “Information” category	191
Figure 53. Grades assigned to solutions from each participant in the “Promptness” category	192
Figure 54. Grades assigned to solutions from each participant in the "Resonance" category	194
Figure 55. Grades assigned to solutions from each participant in the "Inclusion" category	195
Figure 56. Grades assigned to solutions from each participant in "Equality"	196
Figure 57. Grades assigned to solutions from each participant in the "Aesthetics" category	197
Figure 58. Grades assigned to solutions from each participant in the 'Novelty' category	198
Figure 59. Chart presenting character ranking	199
Figure 60. Three papersheet work resulting from workshop subgroup’s exercise	205
Figure 61. Key points of our research within the VSD phases in which they were planned.....	209
Figure 62. Summary of benefits and harms from the stakeholder presentation at the Zurich workshop	211

Figure 63. Comparative Diagram Illustrating the Evolution of Values and Norms	213
Figure 64. Verifier main screen workflow created by a collaboration between Haline Maia and Priscila Ariel (Globo)	234
Figure 65. Configuration workflow allowing users to customize the verifier before accessing the main application screen	235
Figure 66. System workflow for our verifier synthetic generation proposal	235
Figure 67. Screen Flow about how our users will experience the platform on mobile.....	238
Figure 68. Distribution of Responses in AI Ethics, Fact-checking, Artificial Intelligence, and Health Misinformation	248

LIST OF TABLES

Table 1. Responses to the disinfodemic problem divided by four umbrellas.....	40
Table 2. Recommendations for Dealing with moral Dilemmas	55
Table 3. Quality Assessment Checklist.....	99
Table 4. Fact-checking technologies divided by categories, references and quantity.....	101
Table 5. Code framework with examples of the interview script	123
Table 6. Summary of the research used	131
Table 7. Highlights from the semi-structured interviews.....	146
Table 8. Written feedback gathered in the interview about aesthetic.....	148
Table 9. Codes associated with our research questions and asserts from Interview	149
Table 10. Written feedback in the interview with Priscila Arial about technical challenges	150
Table 11. Technological Probes presented at the first Workshops.....	155
Table 12. Participant names and details	157
Table 13. Values and norms presented by each group.....	160
Table 14. Participant names and codes with details.....	162
Table 15. Benefits and Harms for the solutions from each group	166
Table 16. Asserts from group 3 discussions about harms and benefits	167
Table 17. Values and norms presented by each group.....	168
Table 18. Participant names and codes with details.....	171
Table 19. Values and norms presented by each group.....	172
Table 20. Insights about diversity and ethics	173
Table 21. Asserts concerning data and promptness.....	174
Table 22. Demographic Profile and identifications of Study Participants	176
Table 23. Technological Probes presented at the Educage Workshop	178
Table 24. Descriptive Statistics	183
Table 25. Highlights on first impressions of solutions by theme	202
Table 26. Highlights of caregiver experiences by theme	203
Table 27. Highlights on experiences with innovations.....	203
Table 28. Group 1 participants explain their choices.....	206
Table 29. Group 2 participants explain their choices.....	206
Table 30. Group 3 participants explain their choices.....	207
Table 31. Asserts from co-design workshop# 1	215
Table 32. Asserts from co-design from group 2.....	216
Table 33. Asserts from co-design group 3.....	216
Table 34. Asserts from co-design workshop# 2.....	217
Table 35. Asserts from co-design workshop# 3.....	218
Table 36. Author's Ethical Matrix into Fact-Checking Disclosure Research Insights	222

Access to reliable and accurate information is critical at the best of times, but during a crisis such as the ongoing COVID-19 pandemic, it can be a matter of life and death. (Posetti and Bontcheva 2020)

CHAPTER 1: Introduction

The topic of disinformation is not new in either Computer Science or Communication Theories. Still, long before it became a prominent subject in cyberculture studies or engineering, disinformation was a strategy of power assessed by Machiavelli (De Caprio and Salvo Rossi 2022) as a means of maintaining governments or as a practical objective adopted in history, such as Joseph Goebbels' creation of the Ministry of Public Enlightenment and Propaganda to disseminate hate messages against Jews using various media (M.D. Jones 2021), including the press. A funny case in the realm of “health misinformation” dates to 2 June 1897, when reports from New York claimed that Mark Twain was “dying in poverty in London.” This was fake news, as confirmed by Mark Twain himself (Hill 2010). Additionally, one cannot forget the valuable role of the press during past pandemics since the fact that the name “Spanish flu” does not refer to the disease's country origin but to the fact that the Spanish press, which was free during World War I, was known for reporting on it globally while other countries suppressed the news to avoid affecting soldiers' morale.

Despite its long media history struggles and applications in research, the last few years have made the study of this topic more relevant than ever. On 11 March 2020, the day the COVID-19 pandemic was officially declared by United Nations, an overwhelming surge of information — termed an infodemic — flooded the internet and social media platforms. This deluge included accurate updates as well as widespread misinformation. The rapid spread of false information exacerbated public confusion and fear, complicating effective health crisis management efforts.

Technologies and social media networks played a dual role; while they facilitated the dissemination of crucial health information and guidelines, they also amplified misleading and harmful content. The challenges of verifying facts and curbing the spread of false information highlighted significant issues regarding using Artificial Intelligence and regulating digital platforms during such critical times. It was within this challenging context that our research began in 2020.

1. Problem and motivation

This research will study whether technologies used to spread health misinformation can help prevent it. To investigate this hypothesis, we will explore the available health infodemic data and the journalistic coverage of the most recent pandemic. Infodemic, the joining of the words information and pandemic, defines the rapid and far-reaching spread of inaccurate information (2020, 67). According to Posetti and Bontcheva (2021), access to information allows the right to freedom of expression to be realized, which is important and useful for citizens. Verifiable and reliable information produced by professional journalism is vital to achieving "Knowledge Societies" (Posetti and Bontcheva 2021).

Considering that the combination of AI and disinformation corresponds to a broad area of research, we have narrowed our study to the context of health information - a well-defined and contemporary problem. The study focuses on misinformation, defined as false information shared without harmful intent (Canavilhas and Jorge 2022), differing from disinformation, which involves deliberate deception (Lazer et al. 2018).

N.J. Nilsson (1998) defines Artificial Intelligence as the machine's ability to reproduce human capabilities, such as search, planning, knowledge representation, robotics, natural language programming (NLP), and learning. These applications also correspond to uses within goals related to journalism, such as synthetic media, a term for the artificial production of media by automated means, usually through artificial intelligence algorithms based on data (Schick and Schütze 2020). AI journalistic initiatives are also included in fact-checking initiatives and include

datasets analysis (Hassan, Li, and Tremayne 2015) claim reviews (Saeed et al. 2022), and evaluation campaigns (Barrón-Cedeño et al. 2020)

One of the bottlenecks in dealing with fake news is the speed at which it is spread compared to the longer time it takes to verify the news and publish its results. According to Vosoughi, Roy, and Aral (2018) false content is more likely to go viral compared with real. They point out that a false story reached readers six times faster than a real story in a study based on a dataset of rumors from Twitter from 2006 to 2017. For infodemic problems, technologies to help with fact-checking have been explored extensively in studies (Atanasova et al. 2019; Baly et al. 2020; Coello-Guilarte et al. 2019; Gencheva et al. 2017; Hassan, Li, and Tremayne 2015; Nakov et al. 2021; Nguyen et al. 2020; Oshikawa, Qian, and Wang 2020; Pogorelov et al. 2020; Zuo, Karakas, and Banerjee 2018). What needs to be explored more is how fact-checking initiatives could facilitate access to their verified content since even with the many technological studies that have been done to help in the process, the discrepancies with the velocity by which fake news spreads remain (Perloff 1993; Lazer et al. 2018; Galhardi et al. 2020).

An initial approach to the problem will be to understand the impact of the use of AI in synthetic media, which has become popular in controversial uses such as deepfake. Synthetic media refers to the broad category of media that is artificially generated or altered by machines, particularly through the use of artificial intelligence and machine learning technologies (Millière 2022). Synthetic media is also known as generative media (Bateman 2022), and is a type of virtual, often realistic, media created using AI with deep learning techniques. Deepfakes are a subset of synthetic media (AI-generated videos, images, sounds, and text). It is possible to generate deepfakes - realistic video or audio content of virtually anyone saying or doing anything (Hwang, Ryu, and Jeong 2021). This type of content can induce unshakeable beliefs in facts that never happened. Moreover, AI systems can generate millions of fake identities - so-called bot armies, capable of rapidly producing billions of comments, tweets, and recommendations daily, burying journalistic efforts to disseminate truthful information (Stuart Russell 2022).

We recognize the challenge of working on such a broad topic. However, the interest in studying new technologies in depth became necessary in the face of the

need for more women to research in the field. Even so, the aggravating factors of the COVID-19 pandemic meant that in addition to studying AI, the area of health misinformation was crucial to being anchored in this utilization. Thus, facing this challenge had a component of seeking relevance and a social legacy to leave with our contribution inherent to the very days in which this research was carried out. Therefore, the methodological construction of this research will aim to narrow the focus of the problem to be addressed without losing the plurality of listening to various stakeholders.

2. Objectives of the research

The research's primary objective is to understand the use of Artificial Intelligence in journalistic initiatives supporting the citizens under health disinformation. Figure 1 shows how our research proposal aims to observe the ecosystem of fact-checking technologies for health misinformation. The current moment has brought possibilities using synthetic media and automatic journalism that will be studied and considered.

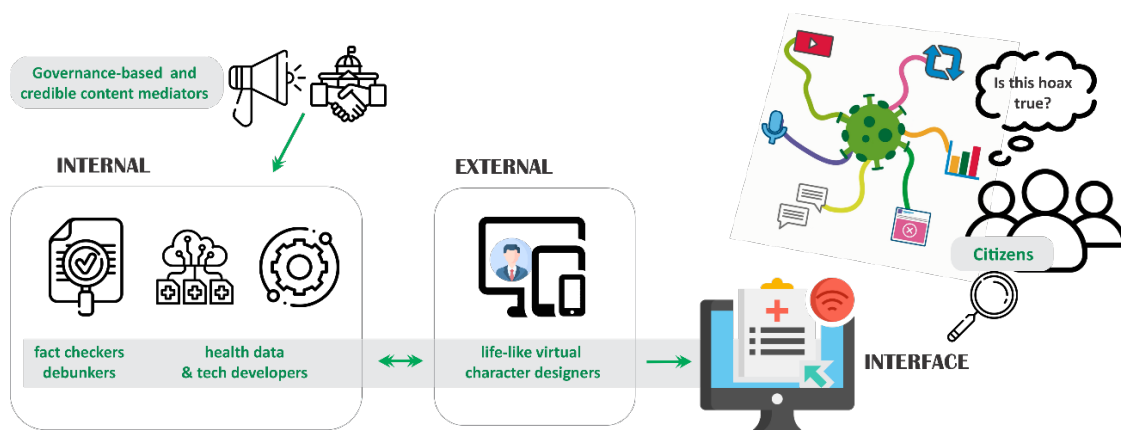


Figure 1. Intervention ecosystem proposal inspired by UNESCO Disinfodemic study

The secondary objectives rely on exploring specific techniques to provide risk perception and give citizens the information to govern themselves. The first is to research fact-checking, which seeks to verify factual information and the accuracy of the statements (Miller 2020). The secondary objective is to understand if the new age of artificial intelligence for journalism addresses the journalistic core values

correctly and ethically. The research also points to the possibility of bridging the gap between the fact-checking organization's content and the public understanding of its role with synthetic media summarizing the most relevant ones (Kickbusch 1997). As a third objective of this study, and using the input collected from stakeholders interviews, we intend to conceptualize the design of a software system that provides health alerts (Felle 2016).

3. Hypothesis and Research Questions

The hypothesis to be analyzed is:

H: AI-powered communication tools have the potential to amplify the effectiveness of fact-checking efforts against health misinformation, promoting informed decision-making. Nevertheless, the use of these technologies can be carried out ethically.

To analyze these assumptions, we believe that we will have to answer four research questions, each focusing on the areas that relate to this hypothesis, which are misinformation in health, artificial intelligence, fact-checking, and AI ethics. The first research question to be answered is as follows:

RQ1 - Health Misinformation: In what ways can the integration of artificial intelligence in disseminating fact-checking outcomes help mitigate health-related misinformation on digital platforms?

To answer this, we will examine the current state-of-the-art in health fact-checking, integrating AI and automation, and exploring their potential applications. This involves participatory research and targeted interviews to gather firsthand experiences and data from workshops.

- a. Which AI technologies have demonstrated effectiveness in fact-checking during the pandemic, and how can a systematic review of these technologies contribute to identifying themes for our research?
- b. What digital initiatives are actively addressing health-related fact-checking and employing innovative methodologies for health specifically?

Additional Questions:

RQ2 – Artificial Intelligence: How can artificial intelligence facilitate disseminating fact-checked health information?

This question requires exploring synthetic media, data journalism, and automated journalism. We will investigate:

- a. The limitations of Synthetic Media using AI in health fact-checking.
- b. Public perception of Synthetic Media and deepfake technology in health fact-checking.
- c. The role of automation speed in curbing misinformation during health crises.

RQ3 - Fact-checking: In what ways can the process of health fact-checking benefit from the incorporation of artificial intelligence to speed up the dissemination of verified information on digital platforms?

This includes researching details about fact-checking processes and initiatives. We will explore:

- a. The effectiveness of chatbots in fact-checking health misinformation.
- b. Initiatives that provide summarized dissemination and automation on digital platforms for health hoax awareness.
- c. The potential applications of synthetic media created using deep learning for fact-checking purposes.

RQ4 - Ai-Ethics: How can we design AI-driven solutions to disseminate health-related fact-checking, ensuring ethical principles as a fundamental requirement?

In this question we want to investigate the potential for a symbiotic relationship between fact-checking organizations and AI that keep journalistic integrity and ethical standards. This research will explore:

- a. Ethical frameworks and guidelines that could be used for innovative fact-checking dissemination while maintaining ethical integrity.
- b. Use human values as a criterion for evaluating our prototype's framework.
- c. Strategies for ensuring inclusivity, representativeness, and adherence to ethical values in automated fact-checking tools.

4. Narrowing the topic: Technology Applied to Health Fact-Checking

Despite growing research interest in synthetic media studies, it is surprising that there is still a substantial gap in research concerning public reception leveraging synthesized media in health fact-checking, examining its effectiveness, ethical implications, and audience engagement.

We propose to build on Coddington's (2015, p. 6) concept of “practices or services built around computational tools in the service of journalistic ends”, exploring how these tools can be tailored for health-related news. A critical area underexplored is the perception of audiences towards AI technologies in health news, particularly in scenarios where misinformation can have dire consequences. Understanding this perception is key to developing AI applications that are both informative and trustworthy.

This research will part from an innovative approach that crisscrosses proven methods. It combines qualitative and quantitative techniques using systematic review, focus groups, interviews, co-design workshops, and surveys to gather comprehensive data on these matters. The methodology aims to gain an in-depth understanding of public attitudes and interactions with AI-based health fact-checking. By doing so, the research intends to contribute to a process that values truth and actively promotes a better-informed stakeholder, potentially reshaping the public's understanding and consumption of health news.

We also know that misinformation is a complex problem, and we intend to narrow down the subject with technologies used to fight health misinformation to obtain better research results. Wicked problems, as discussed by Rittel and Webber (1973), are complex and socially constructed, involving conflicting interests and requiring a policy mix. These problems cannot be immediately tested for results, making their resolution challenging. Addressing multifaceted dilemmas involves networks and collaborations across various stakeholder domains, particularly when tackling societal challenges and driving socio-technical transitions toward sustainability. This involves legitimizing justice and other innovative practices (Hoffman et al. 2021).

5. Overview of the Thesis Structure

This research is organized into four focused chapters that examine how social actors engage with contemporary society, where advancements in AI are driving new communication methods and challenging existing structures, with an emphasis on using AI for disseminating health-related fact-checking. The first chapter lays the ontological foundations, expressed through paradoxes present in the new organization of the media and technological ecosystem. It is divided into two parts. The first part, titled "From Digital Ecosystem to Health Disinformation and Fact-Checking," includes a literature review, beginning with an exploration of the current digital ecosystem and the role of technology in shaping new communication forms. This is followed by a discussion on post-truth and how mediation processes affect public perception and communication. Further, it analyses the phenomenon of disinformation, particularly in health, and its various challenges. The appearance of fact-checking in maintaining journalistic integrity and accountability is also highlighted.

The second part, "Synthetic Media Contribution to Share Health Fact-Checking," delves into the contributions of generative media in disseminating fact-checked information. It examines the ethical considerations surrounding the use of AI in media, discusses the automation processes and the role of AI in generating media content, and explores the potential and challenges of automating the fact-checking process. Concerning the use of automation in fact-checking, this phase of our research was conducted adopting a PRISMA systematic review that served to offer meta-analyses that, in some way, helped to direct the third chapter of our investigation.

The second chapter details the methodological approaches used in this study, framed within a Value Sensitive Design (VSD) framework. It begins explaining our conceptual investigations, laying the theoretical groundwork relevant to the study. It then moves on to technical investigations, examining the technical aspects and capabilities of the tools and methods to guide our research. The empirical investigations phase focuses on data collection and analysis. This

chapter also identifies and discusses the various stakeholders participating in the research and highlights the challenges encountered during the research and the strategies employed to mitigate them.

The third chapter presents the technical and empirical findings of the research. It begins with an examination of case studies and interviews related to the publication of fact-checking results, followed by an exploration of the use of virtual characters and deep learning in disseminating information. The development and execution of co-design workshops to gather insights from various stakeholders are detailed, along with the scenarios used to assess the values involved in the research. This chapter also examines the application of VSD methods in a diverse cultural setting and focuses on a specific workshop aimed at understanding value-oriented approaches in disseminating health information to the elderly.

The fourth chapter integrates the values, norms, and technical requirements identified throughout the research. It discusses how values can be integrated into the design process of technological tools and outlines the ethical guidelines and principles that should govern the design and use of AI in media. A mock-up proposal for a value-driven technological solution is presented, along with a summary of the findings and answers to the research questions posed at the beginning. This chapter also discusses the limitations of the proposed framework and areas for future research.

In conclusion, this research navigates the complexities of integrating artificial intelligence in combating health misinformation, balancing ethical considerations, and leveraging technology to foster new communication paradigms. The exploration spans theoretical foundations, empirical research, and practical applications, culminating in a comprehensive analysis that bridges the gap between technology and societal needs.

The pandemic's effects on the everyday lives of people are profound and long-lasting, spanning all dimensions of human existence: biological, political, economic, technological, cultural, and social.

(Borin et al. 2024)

CHAPTER 2: Review of the State of the art

Section 1: From digital ecosystem to health disinformation and fact-checking

This chapter addresses two primary perspectives. First, it analyzes cyberculture theories, beginning with the digital revolution that has led to the emergence of digital journalism and AI-powered communication technologies. It examines the differing concerns regarding the Internet's impact on society, highlighting its social, economic, and communicational implications. Cyberculture theories reflect a longstanding debate in communication studies: the moral implications and the effects of technology on individuals and society.

Second, the chapter explores the concept of post-truth through various authors and methodologies, distinguishing it from the notion of fake news. This distinction includes understanding different epistemologies related to disinformation and the role of fact-checking within this context. The chapter concludes by presenting the challenges related to health misinformation within this framework, employing a deductive literature review that moves from general concepts to specific consequences.

1.1 Digital ecosystem and technology

The thesis addresses two main areas to understand how objectives and social actors can leverage technology to demand new forms of communication and challenge established structures. The first part focuses on the digital revolution in

journalism, exploring its ontological foundations and the paradoxes within the media and technology ecosystem, particularly in the context of fact-checking and post-truth. It also examines the implications for health communication.

Technological innovation has always driven significant changes in journalistic communication. From the transition from manuscripts to printed press facilitated by Gutenberg's movable type to the influence of radio and television, journalism has continuously evolved with technological advancements. The Internet and AI have further transformed journalism, challenging traditional print newspapers and introducing a new media ecology that affects all stages of journalistic activity, from content production to distribution and consumption.

Digitization plays a critical role in understanding data and correcting errors. The rapid evolution of technology, including robotics and automation, necessitates a deep understanding of how health literacy intersects with these advancements (Bélanger and Carter 2008). The media ecosystem, built on digital data, enables new ways of understanding and disseminating knowledge. Technology decentralizes information, fostering interactivity and connectivity among individuals (Darley 2009).

Digital media philosophy posits that digitization is conditioned by, but not solely determined by technology. Human programmers impart inherent and often invisible characteristics into digital systems (E.C. Silva 2013). This perspective allows for greater pluralism and challenges the close ties between journalists and elites, such as those in public health policy. Younger generations, native to the technological context, are particularly adept at navigating this new digital landscape.

However, technology's accessibility does not always equate to democratization. Health literacy has been a frontrunner in utilizing available data due to its attractiveness. AI and robotics in health promotion, such as apps for disease monitoring, nutrition, or physical activity, are examples of impactful technological use. E-Health, defined as using the Internet and telecommunications in healthcare, exemplifies this integration (Eysenbach and Till 2001).

The practicality and instrumentality of technology are evident in how society organizes knowledge and uses technology. Within journalism, AI has a significant

social component. Simondon's work helps to understand the complexities of technology, tools, and machinery, emphasizing their purpose and contribution (Simondon 1989). Digital tools and applications in health information dissemination demonstrate the potential of digital life.

Techno-realism critiques the economic and exclusivist controls on technology, highlighting how power uses futuristic theories to obscure real problems (Barbrook 2009). Economic context directly influences access to information technology, with more affluent populations having greater access (Cohn and Cope 2011). This dynamic underscores the disparities in technological reach and utilization. The following section will explore the cyberculture considerations surrounding digital innovation and their criticism.

1.1.2 Cyberculture theories and critics

Cyberculture theorists have defended the Internet's moral, political, and economic virtues, particularly professionals linked to communication businesses. Rüdiger (2016) provides a helpful chronology of Internet, categorizing them into three principal evaluations: populist, conservative, and critical. Though potentially pejorative, this classification extends beyond mere business technology and communication articulations.

Rüdiger (2016, p. 27) traces this aspect of cyberculture back to Marshall McLuhan, who viewed the computer society in a utopian manner. McLuhan's assertion that "the medium is the message" emphasizes media's social impacts on organization, independent of their content. For instance, the significance of the telephone lies not in the conversations it carries, but in the social, geographical, and urban changes it instigates (McLuhan 1971, 104).

In this context, technology can be seen as a fundamental raw material, akin to energy sources. McLuhan (1969, 37) illustrates this by comparing media like radio and television to essential commodities like cotton and oil, which profoundly influence the cultural fabric of society.

Pierre Lévy aligns with this perspective, viewing societal changes through the lens of technology without addressing its control mechanisms. Consentino (2020)

notes that Lévy (1997), by discussing technological changes without considering the power of capitalism, affiliates with the technophile movement. Lévy's optimism, however, seems rooted in a new Enlightenment project of knowledge rather than an uncritical enthusiasm. In "Cyberculture," Lévy acknowledges:

Nevertheless, despite my optimism, I do not believe the internet will magically resolve all of the planet's social and cultural problems. I do, however, want to acknowledge the following: First, the growth of cyberspace is the result of an international movement of young people eager to experiment collectively with forms of communication other than those provided by traditional media. Second, a new communications space is now accessible, and it is up to us to exploit its most positive potential on an economic, political, cultural, and human level (Lévy 1997, IX).

Lévy's concept of collective intelligence presents a promising vision of digitization. He recognizes its potential for isolation, dependence, and exploitation but emphasizes its participatory and socializing capabilities. This perspective highlights the divide between the info-rich and the info-poor, exacerbated by digital technology. At the heart of Lévy's positive reflection on the Internet is the understanding of collective intelligence, which he associates with three dimensions: power-sharing (cyber-democracy), productivity and prosperity (information capitalism), and spiritual and artistic enrichment (virtual worlds and games) (Lévy 1998, 192). These dimensions underscore real-time coordination, a hallmark of collective intelligence distinct from past human cooperation.

Castells (2003) further explores the Internet as a new public space, particularly in social networks, which fragment and amplify voices, enhancing social autonomy. This networked space, merging digital and physical realms, fosters autonomous communication and social movements, connecting individuals beyond the control of traditional power structures. Although Castells (2005) later adopted a more pessimistic view of democracy's future, his earlier work remains relevant for understanding the intersection of technology and social autonomy.

As described by Rüdiger (2016, 26), cybercriticism examines the connections between cyberculture and power (political, social, and economic), considering the challenges it poses for individuals. Cybercritics argue that public participation driven by market forces transforms individuals into products. (Siegel 2008) contends

that focusing on audience approval leads to a logic dictated by corporate algorithms rather than spontaneity.

Jenkins (2011) discusses convergence as the merging of old and new media, emphasizing content flow across platforms and media industries' cooperative behavior. He highlights the cultural shift towards collective intelligence and participatory culture, noting the complexities and uncertainties of this transition:

Contradictions, confusion, and multiple perspectives should be anticipated at a moment of transition where one media paradigm is dying, and another is being born. None of us knows how to live in this era of media convergence, collective intelligence, and participatory culture. These changes produce anxieties and uncertainties, even panic, as people imagine a world without gatekeepers and live with the reality of expanding corporate media power (Jenkins 2011, 170).

The new online media ecosystem has led to the rise of new players in the gatekeeping process, increasingly marked by a dynamic of disintermediation. While this shift allows a growing number of individuals and groups to gain visibility, voice, and agency — overcoming contexts of social isolation or quarantine through alternative means — it also means that many accessible contents can now be consumed as information without the same level of trust and accuracy as before (Amaral and Santos 2019).

Other theorists, such as Burke and Briggs (2002, p. 266), highlight convergence's role in blending media and telecommunications, demonstrating how technology can dissolve ideological boundaries. Latour's actor-network theory, applied to cyberculture, views tools and technology as equal agents in social action, reshaping our daily behaviors (Lemos 2013). Other more recent works look at the relationship with technology and communication through it as a state of resonance and alienation. This is the case with Hartmut Rosa (2019), who, in his work "Resonance: A Sociology of Our Relationship to the World," argues that modern life's accelerating pace has not increased happiness. Instead, achieving resonance—harmonious engagement with the world — is essential. Rosa contends that our quality of life depends on our ability to resonate with our environment, a challenge intensified by modernity's relentless escalation logic (Rosa 2019, 202).

Rosa's theory suggests that understanding our fractured relationship with the world through the concept of resonance offers insights into the crises of modern

society. Empathy and mutual understanding are crucial for addressing these crises and fostering a more connected and meaningful existence:

The notion that linking of our physical, affective, and cognitive relationships to world can be understood as a result or process of mirroring – or, rather, of resonance – not only is the underlying idea of this book, but also turns out to be an astonishing point of convergence of disciplines. (...) The recent rise of empathy research as a key topic in cognitive science doubtlessly provided a starting point for this development, having shown that the capacity for mutual empathy, understanding, and adopting other perspectives is important not just for the development of sociomoral qualities but for learning, thinking, and acting in general (Rosa 2019, 145).

The concept of resonance is explored through research analyzing the production of memes, which facilitate moments of connection in a world characterized by alienation. Moreno-Almeida (2024) conceptualizes memes as "digital monsters" that aim to create spheres of resonance, sometimes excessively, enabling a moment of connection between individuals in a predominantly alienated society. In this sense, the author describes an interesting concept of the digital grotesque:

Memes thrive on hybridity while also relishing their ability to return. Their condition as undead guarantees their revival. Memes, as multimodal digital artifacts, flourish in chaos and disorder, blending written words with images (moving or still) and sounds. They appear as a main artefact of political commentary as well as reactions on social media platforms. Memes foster transmedia online participation (Jenkins 2006, p. 2016), hijacking stories, diverting and multiplying them, and ultimately adding layers of meaning uncontrollably. Memes cross borders of medium and challenge traditional storytelling and notions of correctness, often flourishing by reflecting concealed aspects of society (Moreno-Almeida 2024, 35).

This understanding is aligned with Shifman (2013, 41) definition of memes as digital items with shared characteristics, created, circulated, and transformed with mutual awareness among users. Such interactions form a complex web of digital engagement that influences civic culture and political discourse, particularly during crises. Resonance is crucial for comprehending how memes achieve widespread circulation and impact.

This concept goes beyond virality, highlighting the depth of connection and shared understanding that memes foster among users. Scholars contend that meme resonance is attained through humor, irony, and cultural references, collectively creating a sense of solidarity and shared identity among viewers (Phillips and Milner

2017). The resonance effect of memes is evident in their ability to convey complex social and political critiques in a relatable and often playful manner, thus influencing public opinion and discourse. This phenomenon is particularly observable in the proliferation of pandemic-related memes commenting on governmental strategies and public health measures (Doona 2024). By generating resonant content, memes contribute to ongoing dialogues within digital spaces, underscoring their role as potent agents of cultural and political expression, affective reactions and persuasive potential (Lundqvist 2019; Geniole et al. 2022; Kahan et al. 2017; Wasike 2022).

During the COVID-19 pandemic, memes have played a critical role in reflecting and shaping public trust and vulnerability toward governmental strategies. According to Roslyng and Larsen (2021) trust entails accepting a vulnerable position relative to a trusted entity. Memes have been employed to critique, defend, and jest about pandemic strategies, often through irony and dark humor. This is evident in the use of dark humor to “articulate the impact of grief and ascribe meaning to loss (Murray 2016, 55).” Other authors are still trying to understand the power of using memes to spread prejudice and xenophobic sentiments.

Furthermore, the interplay of humor and political discourse is significant in understanding the impact of memes. Humor offers an “alien perspective” (Critchley 2011) engaging users in critical reflection and acting as agents of globalization (Shifman, 2014). The political and social critique embedded in memes illuminates intersectional hierarchies and ideological conflicts, particularly within far-right digital landscapes (Critchley 2011). Consequently, memes serve as sources of entertainment and expression and tools for community building, critique, and navigating social and political crises.

This evolving field continues to generate significant scholarly debate and requires a nuanced understanding of its various forms and implications. Authors like Dora D.S. Silva (2021) highlight the challenges of conceptualising journalism within the framework of digital environments. Digital journalism, or journalism in a digital environment, encompasses the production and dissemination of journalistic content specifically for digital platforms. The journalism produced for

and within a digital environment has been referred to by many names, reflecting the rich academic discussion on terms such as "cyberjournalism," "online journalism," "webjournalism," and "digital journalism." Therefore, we emphasise this definition:

(..) we use the expression "digital journalism" as we believe the other terms can be misleading — for example, online journalism, webjournalism or cyberjournalism can be perceived as just referring to journalism practiced and disseminated on the Web, and digital journalism literally includes all the journalism produced and disseminated through digital media platforms, including mobile phones and tablets, for example (which can deliver offline as well as online content) - (D.S. Silva 2021, 76).

Efforts to drive innovation within media have been varied, ranging from highly successful to less effective initiatives. Key areas of focus include the introduction and enhancement of products, the refinement of processes, the redefinition of organizational positioning, and the transformation of organizational paradigms (Francis and Bessant 2005). According to Granado, Silva, and Vicente (2020), in the media and creative sectors, this might involve the development of new applications, the implementation of novel storytelling techniques such as gamification and augmented reality, the repositioning of organizations with a focus on digital content, or the overhaul of business models. Contemporary research has further expanded to examine journalistic genres and formats, the interplay between innovation and crises, social innovation, and the importance of open innovation (Granado, Silva, and Vicente 2020).

The rapid advancement of digital technologies and their increasing convergence have enabled innovation in the early 21st century to become disruptive, with the potential to revolutionize entire industries rather than merely introduce incremental changes (Granado, Silva, and Vicente 2020). This notion of transformative change is echoed by scholars such as Pavlik et al. (2018), who contend that the media industry is undergoing significant disruption, characterized by its complex and change-driven nature.

In the subsequent section, we will delve into AI tools for detecting misinformation in the digital age, examining their relevance in the context of post-truth mediation and fact-checking. This discussion builds on the previous section's exploration of AI applications in journalism.

1.2 Post-truth and mediation

These analyses of cyberculture are essential for understanding the relevance of the proposed solutions. The differences between them pertain to the impacts of the Internet on society and its social, economic, and political implications. In this brief overview, cyberculture theories reflect a longstanding debate in communication theories: the moral debate and the discussion about the effects of various aspects of communication on individuals and society. Based on a multifaceted and not always convergent vision of cyberculture, we present the concept of post-truth. Although post-truth is not a central concept of this thesis, its relationship with new information technologies is significant.

Post-truth conceptualization, aligned with cyberculture approaches, may have originated in different cultural contexts and critical strands. As Rüdiger (2016) defines, post-truth is generally more articulated with conservative thinkers or cybercriticism. The current reflections indicate an apparent movement in the debates within the field of Communication: at least in the theoretical-reflexive discourse, as opposed to the dominant narratives of technology companies or a significant part of academic reflections on media during the hegemony of traditional offline vehicles, these are not times of optimism or belief in progress.

A significant milestone for the critical questioning of digital media was the inclusion of "post-truth" in the Oxford English Dictionary in 2016 when the term was declared the word of the year. It is described as "relating to circumstances in which people respond more to feelings and beliefs than to facts." Since its inclusion in the dictionary, a more intense debate has emerged in academia about the term's relevance. Some authors argue that post-truth expresses a phenomenon that has always existed in politics and public opinion formation Keyes (2004) offers an intriguing perspective: post-truth is not simply a binary opposition between lies and truth. Instead, post-truth represents a third category, distinct from lies or truth, existing in an ambiguous ethical zone that does not directly oppose facts. It begins with people who propagate what could be considered truths.

According to Marcondes Filho (2019, 17), we are experiencing a form of domination that has supplanted rational and republican discourse, significantly altering the role of the media, particularly the spoken, written, and televised press, with the advent of social networks as a decisive actor. He identifies three stages in this transformation: 1) The diminishing formation of public opinion through traditional media; 2) The monopoly of virtual spaces; 3) The suppression of the new public space, ultimately paralyzing opposition.

Although the definition of truth is extensively discussed in philosophy, and despite it not being the central theme of this thesis, Arendt (1967) presents relevant arguments concerning facts, which can help us understand the meanings of objectivity and factual truth in journalism. Arendt (1967, 10) posits that "conceptually, we may call truth what we cannot change; metaphorically, it is the ground on which we stand, and the sky stretches above us." However, this conception of truth is not free from critique regarding the "totalitarianism of truth," a metaphysical perspective. Her contribution to the definition of factual truth acknowledges this concern:

Factual truth is no more self-evident than opinion, which may be why opinion-holders find it relatively easy to discredit factual truth as just another opinion. Factual evidence, moreover, is established through testimony by eyewitnesses – notoriously unreliable – and by records, documents, and monuments, all of which can be suspected as forgeries. In the event of a dispute, only other witnesses but no third and higher instance can be invoked, and settlement is usually arrived at by way of a majority; that is, in the same way as the settlement of opinion disputes – a wholly unsatisfactory procedure, since there is nothing to prevent a majority of witnesses from being false witnesses. On the contrary, under certain circumstances, belonging to a majority may even encourage false testimony (Arendt, 1967, p. 10).

To align this broad concept, it is essential to understand what is considered objectivity by highlighting the central points of opposition between fact (the substrate of journalistic production) and lies. Sodr  (2019) discusses the tradition of truth from a different philosophical position. He suggests that there is an effort towards objectivity, of Kantian origin, in which an object is the same for all subjects and their experiences. Science, with its hypotheses and theories, is structured under this premise. In this perspective, the world of facts is based on practical experience. This reasoning is fundamental to the construction of Comte's positivism, to which journalistic tradition is an heir:

Journalism incorporates common sense about facts but mainly a sense shaped by positivism, a doctrine whose heyday coincides with the prestigious rise of the European bourgeoisie. Nevertheless, positivism is reinterpreted by the press with the nuances of the morality of history and progress, linking the intelligence of facts and causes to the so-called 'objective' information. (...) The journalistic information starts from objects primarily considered factual to obtain, through the event, some clarity about the socio-historical fact (Sodré, 2019, p. 94).

Constantino (2020) argues that Nietzsche's understanding should be integrated into this reasoning, as there are no facts, only interpretations. Language is the critique of the concept and the transformation from unequal to equal. In observing the world, the objects aligned in the same vision are neither identical to themselves nor what language seeks to define. Language is not the true expression of the world but a necessarily rhetorical expression, where metaphors otherwise express the object as an initial condition for the concept's existence (Consentino 2020).

In this tradition of truth, even as journalism seeks to differentiate reportage (the objectivity of facts) from opinion (the interpretation of facts), Sodré notes that throughout its history, newspapers, consciously or not, have conveyed worldviews that translated and expressed facts in interpretative ways. We can perceive, dialectically, that even being clear about its importance as a constructor of reliable news, subjectivity is also crucial to interpretation, construction, noticeability, and critical capacity (Sodré, 2019). Understanding the challenges of journalism, in realizing subjectivity within rationality before the internet, involves understanding the digital world as a space for debate.

Habermas (1974) explains that the concept of the public sphere, as a space for debate and the use of argumentative reason in the Greek agora, began to be incorporated into the media as they gained validation. The public sphere, configured in the press, represents the space for debates on issues previously practiced in social spaces such as squares, clubs, and cafes. This public sphere, reduced to "means of education," can no longer be considered a principle of enlightenment or a sphere where reason is realized. The public sphere simply integrates subjective opinion into the objectivity that the spirit has given itself in the form of the state (Habermas 2011, 292).

Regarding the use of reason within the concept of the public sphere, when considering the freedom of expression of diverse worldviews, the internet and new

technological strands appear as a beacon illuminating the sea to save the wastelands abandoned by both state and private interests. Thus, the absolute power of the public sphere in the face of technological revolutions cannot be denied.

However, as we analyze in this compendium of dichotomies, all that is utopian in journalism and the internet, in terms of following and recreating digital and multimedia journalism, must evaluate the potential for manipulation. A fundamental complement to Habermas was elaborated by Lippmann, who treated Public Opinion as a result of massification imposed by those in power. In this environment, what determines the communication game are the precepts indirectly created by the media and the news in our unconscious, making information uniform. These stereotyped images of reality control behavior and determine the polarized mood of the public, an abyss that nullifies the space for critical reflection (Lippmann 2017).

In a new digital age, concepts of deliberative democracy in the public sphere could be incorporated into new activism articulated by networks as a virtual agora, a term attributed to Pierre Lévy to describe the potential of collective intelligence on the Internet. Here, new activism articulated by networks is treated with optimism, particularly emphasizing virtual agoras, referencing the Greek squares where accessible discussions about politics and culture were practiced. Lévy states that "the virtual agoras of democracy help people and groups recognize each other, meet, negotiate, and contract. In this respect, developing tools for orientation and localization in political, social, institutional, and legal complexity seems indispensable to us" (Lévy 1998, 43).

According to Wardle and Derakhshan (2018), an informational disorder arises, which is more intricate than "fake news" implies. However, as da Silva, de Albuquerque, and Veloso (2019) stated, fake news and manipulated news possess the fundamental attributes of news and adhere to the structural norms of journalistic texts. These characteristics include brevity, clarity, informativeness, and an impersonal tone. The authors explain that fake news is based on falsehoods, whereas manipulated news is rooted in the deliberate omission of truth (da Silva, de Albuquerque, & Veloso, 2019, p. 416). As we transition to the next section, it is

crucial to understand how these themes intersect with health science and journalism.

1.2.1 Health science and journalism

This broad scenario challenges the new technological ecosystems within information and digital cyberculture. With many nuances, we have focused on a particularly relevant area today: health-related journalism. Our proposal aims to establish a robust focus on this area, leveraging the data and problems encountered during the recent infodemic associated with the COVID-19 pandemic. While popularising scientific health content may not yield immediate rewards in the medium, it can significantly contribute to research funding.

Given these challenges, this work proposes to conduct a review of the state of the art, contributing to the identification of the role of communication in health promotion and education. It also explores technology's crucial role in data analysis, dynamic digital journalism channels, and information synthesis. Thus, considering the complexities mentioned, we will present the theoretical dimensions that configure a dichotomous perspective on communication within the context of health literacy and its practices in social interaction, information content, applied technology, and mediation.

A relevant characteristic of scientific dissemination is its media repercussions. According to Charaudeau (2008), this becomes complex because these discourses must be linked to a specific domain. They are constantly hybrid discourses requiring attention to their situational conditions of production. As Charaudeau (2008) and investigations by Giering and Souza (2013) highlight, scientific popularization in the media exists at the intersection of at least three discourses: the scientific, the media, and the didactic.

These complexities of science communication are further intertwined in the four stages known as the mode of continuity (Hilgartner 1990): intra-specialist, inter-specialist, pedagogical, and popular (Figure 2). The intra-specialist stage represents articles published in scientific journals. The inter-specialist stage includes interdisciplinary articles, even when presented at meetings between

researchers within the same field of knowledge. The pedagogical stage encompasses textbook information with more complete and consolidated data.

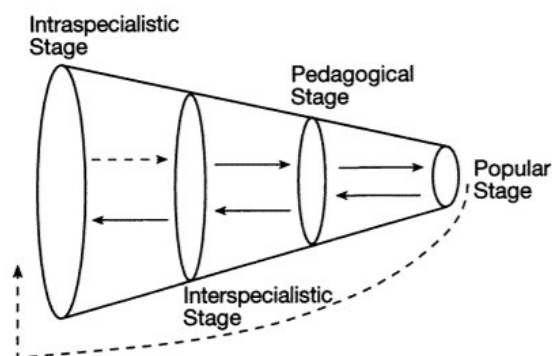


Figure 2. Continuity model of science communication

This interpretation reveals the possibilities for utilizing scientific content that transitions from the intra-specialist expository context to the popular one, passing through intermediate stages. According to Bucchi (1996), the same researcher can communicate his work concurrently at several levels, even if he encounters difficulties in finding suitable forms of communication at any stage or needs to resort to adaptations:

Communication of science at the popular level may influence core scientific practice in many more different and subtle ways than simple support and reinforcement. As already shown, it can foster the inclusion or exclusion of actors or theories from specialist discourse and make room for new interpretations or confer a different status on existing models by linking them to other public issues and themes (Bucchi, 1996, p. 386).

Bucchi (*ibid.*) further notes that the intra-specialist level is foundational. However, reaching the inter-specialist level allows the information to present a broader scientific panorama. When mentioned in other research and books, the theory evolves, and when disseminated through journalistic channels and other mass media, it gains public recognition (Bucchi, 1996, p. 378). The importance of journalism in achieving this popularisation is crucial, especially in health communication. Medicine is particularly prominent in scientific content, a prominence acknowledged by society (Frost, Frank, and Maibach 1997).

According to Arroyave (2012), the increased interest in medical content directly influences the volume of information the media provides. This interest underscores the media's significant role in promoting public health actions and empowering

citizens (Hodgetts et al. 2008). Journalism is often regarded as an organizing force in society (Silverblatt 2004, 39).

Despite its relevance, journalism is undergoing profound changes (Gillárová, Tejkalová, and Láb 2014) within the digital ecosystem. Consequently, health communication systems are also evolving. To understand these transformations, it is essential to grasp the concept of health literacy, which involves efforts to promote collective and individual knowledge and confidence for health decision-making (Kickbusch 1997). Effective communication must create an environment that promotes life through health promotion rather than merely repeating messages.

The World Health Organization (WHO) defined Health Literacy in 1998 as the "set of cognitive and social skills and the ability of individuals to access, understand, and use information to promote and maintain good health" (Nutbeam 2000). According to Nutbeam's (ibid.) definitions, which consider the progressive autonomy of individuals as their cognitive levels evolve, we can correlate with the WHO's levels of Communicative and Critical Literacy. Sørensen et al. (2012) identify critical issues for information management and health: access (the ability to seek and acquire desired information), understanding (the ability to comprehend data), evaluation (the capacity to interpret informational content), and application (maintaining and improving health quality).

These fundamentals validate the primary function of health literacy: enabling individuals to make better health decisions. Evidence-based medicine, a movement strengthened in the 1990s, upholds the principle of ensuring that patients receive the best treatment suitable to their priorities and supported by the best possible evidence (Sackett et al. 1996). Humanistic support organizations further assert that one of the bases for enabling citizen involvement in health care is through decision-making related to disease prevention, treatment, and management (Silverblatt 2004).

An important validation for health literacy arrangements lies in cognitive and psychological factors. McNaughton et al. (2019) found that patients with low health literacy who had suffered acute heart failure were 35% more likely to die within 21 months of hospitalization. Effective communication makes health information accessible and integrated with societal demands.

The term online health information (OHI) informs people about disease prevention, treatments, general well-being, and pandemic risk situations. OHI is traditionally seen in communication theory as a form of enlightenment from health professionals to patients (Paimre, Virkus, and Osula 2023). Currently, the Internet is a major source of health information (Case et al. 2004; Kitchens, Harle, and Li 2014). "Most doctors are already feeling the effects of patients arriving at their offices armed with printouts from the World Wide Web and requesting certain procedures, tests, or medications (Hesse, Greenberg, and Rutten 2016, 769)." This sets the stage for examining the disinfodemic perspective and the associated challenges.

Lima and Sousa (2021) illustrate how journalism has historically used different approaches to portray health crises, as seen in the coverage of the pneumonic flu in Portuguese newspapers during the early 1900s (1918-1919). They highlight the obituary of Dr. António Esmoriz, celebrated both as a compassionate individual and a dedicated physician who succumbed to the very pandemic he fought. Described as a "professional hero," his story underscores the nobility of the medical profession, providing society with inspirational figures. This concept resonates with today's narrative during the COVID-19 pandemic, where healthcare workers are similarly hailed as heroes for their essential roles in combating the crisis. It draws a parallel to how the cultural archetype of the selfless hero, who sacrifices for the greater good, has been a consistent theme in health crisis reporting (Lima and Sousa 2021, 211).

1.3 Disinfodemic perspective and challenges

Understanding how journalism adopts new technologies is fundamental, particularly given the dangers of communication in its negative aspects, such as fake news during the COVID-19 pandemic. Recent studies (Zarocostas 2020) indicate that public health information today relies on the same distribution tools traditionally used to disseminate disinformation. The term "infodemic" (Mendoza, Poblete, and Castillo 2010) has been used to describe the perils of misinformation during pandemic scenarios (Van der Haak, Parks, and Castells 2012). It can

negatively accelerate the epidemic by influencing public behavior (Zarocostas 2020) or even have fatal consequences (Posetti and Bontcheva 2020).

It is essential to recognize the significant harm that disinformation can inflict on democracy by undermining trust in health information. It leads to poor decision-making and disrupts public debate on critical issues; therefore, automated fact-checking must be complemented by highly trained human oversight. This thesis aims to understand the importance of combating health disinformation through technology. Beyond understanding the ecosystem of this issue, it is necessary to grasp concepts relevant to the theme. A reasonable definition is essential, given the apparent affinities between fake news and post-truth.

Tandoc Jr, Lim, and Ling (2018) provide such a definition through a literature review of 34 articles published in English from 2003 to 2017. Fake news, ideally opposed to news and journalism committed to the truth, is more complex than common sense suggests. In their review, fake news is classified into six categories: satire, parody, fabrication, manipulation, advertising and public relations, and propaganda.

Rashkin et al. (2017) classify fake news based on the author's intent, identifying three types: a) hoax: persuading the validity of a paranoid report; b) propaganda: leveraging people's mindsets to follow a specific agenda; c) satire: mimicking real news but indicating that it is not to be taken seriously. Rashkin et al. (2017, 2932) explains, "unlike hoaxes and propaganda, satire is intended to be notably different from real news so that audiences will recognize the humorous intent. Hoaxes and satire are more likely to invent stories, while propaganda frequently combines truths, falsehoods, and ambiguities to confound readers."

The Council of Europe classifies fake news into three categories: 1) disinformation – news intended to harm people, groups, organizations, or countries; 2) misinformation – news produced without malicious intent but based on errors or misconceptions; 3) mal-information – news supported by accurate data but edited and disseminated with immoral intent (Canavilhas and Jorge 2022, 234). This conceptualization aligns with the definition by Lazer et al. (2018, 1094), that "fake news outlets lack the news media's editorial norms and processes for ensuring the

accuracy and credibility of information. Fake news overlaps with other information disorders, such as misinformation (false or misleading information) and disinformation (false information purposely spread to deceive people)".

Consentino (2020) notes that while the conceptualization of fake news appears new, it has a long tradition in journalism and political theory. The definition of fake news, like post-truth, is rooted in postmodernity and digital culture, though not necessarily new. Sodr  (2019) describes contemporary phenomena of fake news attributed to the Internet, highlighting the perversity in social media rumors legitimized by anonymity, emotional motivations, and the deinstitutionalization of social networks.

Bucci (2018, 27) acknowledges the participatory nature of social networks, as described by Castells and L vy, but attributes their role to devaluing factual truth in the current era. For him, the problem lies not in the technology, interactions, or informational exchanges but in property issues and the industrial exploitation of users' clicks and wills. Sodr  adds that the need for novelty in content and repressed emotions guide internet exchanges (2019, 99). He highlights uncritical acceptance, narcissistic affective gratification, and the virality of content that is not distinguished as truth or lies.

Although many authors view misinformation as a symptom of the weakening of enlightenment ideals, the issue is more complex than a simple dichotomy of "digital literacy" versus "fragility of thought." Genesini (2018) recognizes the problem of the exponential growth in disseminating real or fake news with digital technologies. Lazer et al. (2018) note that there are few scientific answers to the impact of fake news on people, cautioning that "mediation of much fake news via social media might accentuate its effect because of the implicit endorsement that comes with sharing" (Lazer et al. 2018, 1095).

The conception of post-truth echoes postmodern philosophy, strongly influenced by Nietzsche, and critiques the excesses of relativism while rejecting the more objective tradition of reason. Some authors are skeptical of governmental action or social network platforms in regulating freedom of expression, noting that truth is more untamable than we think. Nietzsche's model is also reflected in Foucault's critique of reason:

Despite the apparent elegance of fact-checking, the science supporting its efficacy is, at best, mixed. This may reflect broader tendencies in collective cognition and structural changes in our society. Individuals tend not to question the credibility of information unless it violates their preconceptions, or they are incentivized to do so. Otherwise, they may accept information uncritically. People also tend to align their beliefs with the values of their community" (Lazer, 2018, p. 1095).

According to the literature, the intent behind creating or sharing inaccurate content differentiates disinformation from misinformation. In the context of the Coronavirus pandemic, responses vary according to the motivations behind each. Education can correct misinformation while stopping the profit from hoaxes can weaken disinformation. However, the impact of fake news, regardless of its purpose, is potentially the same. This work focuses on health-related disinformation that predates COVID-19. Falsehoods diminishing scientific credibility extend from the resurgence of the 'flat earth movement' to climate change denial, generally for political or economic reasons.

Considering the dialectic presented by cyberculture theories, the internet is crucial for health disinformation and information. It facilitates communication, connecting the production of fake and trustworthy news to their reception and engagement. The World Health Organization (WHO) has termed the rising COVID-19 disinformation a "massive infodemic." UNESCO has described specific characteristics of COVID-19 disinformation, calling it a "disinfodemic" (Posetti and Bontcheva 2020).

According to Posetti & Bontcheva (2020), this debate encompasses two major scenarios: the high-speed digital connectivity aiding pandemic management and the dangers of misleading content being shared virally. The report highlights that COVID-19 disinformation endangers the pandemic response by causing citizens to ignore scientific advice, amplifying distrust in governments, and diverting journalists' efforts from proactive reporting to disproving fake news. The disinfodemic is characterized by 1) emotional narratives and memes, 2) fabricated websites and authoritative identities, 3) simulated images and videos, and 4) orchestrated disinformation campaigns (2020, 5). According to research conducted for UNESCO, responses to the COVID-19 infodemic are categorized into ten types (Table 1) under four umbrella categories (Posetti & Bontcheva, 2020, 7).

Table 1. Responses to the disinfodemic problem divided by four umbrellas

UMBRELLAS	RESPONSES TYPES
Identifying disinformation	<ul style="list-style-type: none"> • Monitoring and fact-checking responses • Investigative responses
Producers and distributors	<ul style="list-style-type: none"> • Legislative, pre-legislative, and policy responses • National and international counter-disinformation campaigns
Production and distribution	<ul style="list-style-type: none"> • Curatorial responses • Technical and algorithmic responses • Economic responses
Supporting the target audiences of disinformation	<ul style="list-style-type: none"> • Ethical and normative responses • Educational responses • Empowerment and credibility labeling

Monitoring and fact-checking responses tended to be done by independent fact-checking organizations, news companies, academics, civil society organizations, and partnerships. An example described by Posetti & Bontcheva (2020) was a fact-checking organization in over 70 countries coordinated through the International Fact-Checking Network (IFCN), the World Health Organization, social media platforms, NGOs, governments, and news media. Social media groups such as WhatsApp, Facebook, Google, and Twitter have also enforced donations to fact-checkers and journalism companies to help expose health disinformation during the pandemic. Technical and algorithmic responses using Artificial Intelligence (AI) to detect and limit the spread of disinformation were highlighted by them:

Some news media and fact-checking organizations are using automated tools to help speed up COVID-19 fact-checking and content verification. One tool is Agence France Presse's (AFP) InVID/WeVerify browser plugin. It is assisting in debunking disinformation about COVID-19 in numerous countries (USA, France, India, Portugal, Netherlands, Colombia, Ecuador, Mauritius) and languages (e.g., English, French, Spanish, Dutch, and Portuguese) by many fact-checkers and news organizations (Posetti and Bontcheva 2020, 11).

(Granado, Silva, and Vicente 2020) notes that creating new relationship structures with audiences is crucial in journalism, as its survival hinges on distinguishing news content from other types of content. This is particularly challenging with younger audiences, who appear to be losing interest in news content in favour of various other options (Mihailidis, 2014). Given the rapid changes, technological disruptions, and information overload, it is essential to

consider the role of news content and informative genres. Enhancing news literacy is critical, especially among young people and educators, to ensure that new generations can identify, gather, and produce credible information (Kawashima-Ginsberg, 2014).

As discussed by Cardoso (2023) journalistic and scientific content undergoes prior certification for authenticity before being shared online. This verification process, focused on factual accuracy, takes place during production and precedes distribution. However, once authenticity is negotiated by communication participants, two scenarios may occur: the symbolic value of the content is either recognized and accepted, or it is contested by the individual, who then decides whether to align with the communicated authenticity or challenge it. When authenticity is questioned, participants might alter the content's meaning, thereby creating deliberate falsehoods despite knowing the original (Eco, 2016).

When sharing false news, these two dynamics also apply. If the symbolic value of a false news story is not challenged, misinformation can become viral because the news format is assumed to ensure content authenticity. In this case, individuals spread falsehoods unwittingly, as they are unaware of the true information (Eco 2016). Individualized authenticity does not necessarily lead to informational disorder, nor is it neutral. In modern culture, the authenticity of a communicative object was linked to its historical origin and authorial originality. This dual articulation determined the privilege of one object over another, requiring analysis of both its physical expression and content (Eco 2005).

1.4 Fact-checking as a way for accountability in journalism

One theme that requires more attention in this study is the concept of fact-checking and its role in addressing disinformation within the context of cyberculture (DeFleur 2013). This study defines and explains the fundamental foundations of fact-checking, integrating various related theories. Fact-checking involves a process that seeks to verify factual truth, providing investigative responses beyond questioning the content's credibility (Miller 2020; Posetti and

Bontcheva 2021). According to The Duke Reporters' Lab annual fact-checking census "Reporters Lab" 2022), 378 initiatives were operating in 2022. The report also notes that the number of fact-checking efforts worldwide doubled from 2017 to 2022, with nearly 400 journalists and researchers working on debunking hoaxes in 105 countries.

The origins of fact-checking trace back to the early 20th century, addressing the growing sense that information could be unreliable. Collective intelligence was becoming controversial, and the capacity for making informed, independent judgments was perceived to be eroding (Canavilhas and Ferrari 2018). Fact-checking in journalism encompasses traditional internal procedures for verifying facts before publication and newer approaches that publicly evaluate the truthfulness of statements made by politicians, journalists, and other public figures (Graves and Amazeen 2019). Internal fact-checking emerged as a distinct role in U.S. news magazines during the 1920s and 1930s when the objectivity norm was solidifying among American journalists. Although newspapers typically do not employ dedicated fact-checkers, the term broadly refers to verification routines and a professional commitment to factual accuracy (Graves and Amazeen 2019).

External fact-checking involves publishing evidence-based analyses of the accuracy of political claims, news reports, or health information. Since the early 2000s, numerous organizations dedicated to fact-checking have emerged globally. These organizations may be part of established news outlets, universities, or other civil society sectors. The overarching goals of these practitioners are to help the public become better informed and to promote fact-based public discourse. Research generally concludes that fact-checking can be effective in experimental settings, although the impact of corrections is often limited by motivated reasoning (Graves and Amazeen 2019, 1).

Some authors (Cazetta and Reis 2019; Dobbs 2012; Seaton, Sippitt, and Worthy 2020) consider that the first step in fact-checking is validating politicians' speeches. Dobbs (2012) notes that former US President Ronald Reagan inspired fact-checking. The movement began in the early years of Reagan's mandate, in 1982 when American newspapers started questioning the president's misstatements during

press conferences. Dobbs said these efforts were primarily refutations of the president's statements.

Marwick (2013) identifies other early unofficial fact-checkers, such as muckraking journalists who challenged the claims of patent-medicine companies. The work of Samuel Hopkins Adams and Upton Sinclair are examples of journalists whose efforts led to federal regulations protecting consumers and legitimate medical businesses from hoaxes (Cassedy 1964). McBride and Rosenstiel (2013) argue that the fact-checking movement began later, linked to post-election comments in 1988 by David Broder, a reporter and columnist for *The Washington Post*. The project's coordinator was journalist Brooks Jackson, the first journalist to specialize in fact-checking (McBride and Rosenstiel 2013; Dobbs 2012). Mantzarlis highlights two essential milestones for the growth of fact-checking: the 2009 Pulitzer Prize awarded to PolitiFact for national reporting and the global rise of the term "fake news."

Most fact-checkers come from traditional media, NGOs, and university-linked initiatives today. According to some authors (Graves and Cherubini 2016; Humprecht 2020), there are mainly two fact-checker models: the 'newsroom model' and the 'NGO model.' The newsroom model is affiliated with a communication company and, although a minority, usually has a broader reach (Graves & Cherubini, 2016). The NGO model, in contrast, involves independent fact-checkers, free from the business constraints of traditional media but lacking their loyal audiences (Lamprou et al. 2021).

The Fact-Checking Network (IFCN), part of the Poynter Institute, follows the NGO model and has established a code of ethics for fact-checkers since 2016. The first fact-checking initiative in Brazil was O Truco in 2014, followed by the Lupa agency in 2015 (Cazetta and Reis 2019). Other agencies, such as Projeto Comprova, led by the Brazilian Association of Investigative Journalism (Abraji), support journalists from 40 Brazilian media outlets. Germany's ARD public broadcaster also operates its fact-checking website.

The NGO model includes platforms like Ellinika Hoaxes, funded by Facebook to detect false stories ("Ellinika" 2019), TV programs such as *El Objetivo* (Ruiz and Verdú 2018) and *Observador* in Portugal since 2015 (Canavilhas and Ferrari 2018)

also contribute to fact-checking. According to Cazetta and Reis (2019), other notable initiatives include FactCheck.org (2003), PolitiFact.com, and the Washington Post’s Fact Checker (2007), as well as HealthNewsReview.org (2004).

Another interesting example of validating fact-checking initiatives during the pandemic is the COVID-19 Fact Checkers Dataset (2020), an extensive international collection of over 200 active fact-checking groups dedicated to verifying COVID-19 misinformation. Managed by the Social Media Lab in collaboration with the World Health Organization (WHO), it aims to study the spread of COVID-19 misinformation and map fact-checking efforts worldwide. The dataset enhances public understanding of the COVID-19 fact-checking landscape and provides data-informed insights for policymakers in combating pandemic-related misinformation. The compilation process involved web searches to locate fact-checking groups meeting the COVID-19 criteria, followed by cross-validation, enrichment, and data cleaning to ensure accuracy and standardization.

To expand the dataset, a three-pronged approach was adopted. A custom geo-based visualization identified countries with limited or no records, prompting relevant keyword searches on search engines and social media platforms. Secondly, information was sourced from news and feature articles about fact-checking groups from reputable outlets. Lastly, the dataset was compared with other fact-checking organizations in the COVID Global Misinformation Dashboard (Figure 3) obtained through the Google Fact-Check Tools API.



Figure 3. A Snapshot of the International COVID-19 Fact-Checking Ecosystem

To enhance the repository's comprehensiveness, the Social Media Lab sought input from the public through a Fact-Checkers Suggestion Form, encouraging submissions of new organizations engaged in COVID-19 fact-checking. Additionally, they provided an Error Reporting Form for users to help maintain the dataset's accuracy by notifying them of any updated information or errors related to the fact-checkers listed.

Fact-checking differs from other journalism formats in several ways. One key distinction is highlighted by Lamprou et al., who notes that fact-checking prioritizes accuracy over simply getting the gist of an idea (Lamprou et al. 2021). Furthermore, while traditional news companies aim to avoid inaccuracies, fact-checkers publicize claims they determine are incorrect (Amazeen 2013). According to Amazeen, the fundamental goals of fact-checking are threefold: educating the public, improving political behavior, and enhancing journalism (2013).

Vlachos and Riedel (2014) describe the typical sequence of fact-checking steps: (a) extracting statements to be fact-checked, (b) constructing questions, (c) obtaining evidence from relevant sources, and (d) deriving a verdict from that evidence. This process is often time-consuming and can take hours or days, by which time misleading statements may have spread widely. The American Press Institute defines fact-checking as:

Fact-checkers and fact-checking organizations aim to increase knowledge by re-reporting and researching the purported facts in published/recorded statements made by politicians and anyone whose words impact others' lives and livelihoods. Fact-checkers investigate verifiable facts, and their work is free of partisanship, advocacy, and rhetoric. The goal of fact-checking should be to provide clear and rigorously vetted information to consumers so that they can use the facts to make conscious choices in voting and other essential decisions (Elizabeth 2014).

Graves, Nyhan, and Reifler (2016) relate the increase in modern fact-checkers to the breakdown of traditional factual and objective news. They identify three factors influencing this rise: changes in journalistic routines, technological transformations reducing the gatekeeping role of conventional newsrooms, and the limitations of public debate presence in the media landscape. Consequently, the growth of fact-checking can be seen as a self-regulatory effort promoting ethical journalistic revision (Pickard 2018).

UNESCO presented guidelines for fact-checking in their handbook for Journalism Education and Training (Mantzaris 2018), which include:

1. Finding fact-checkable claims by scouring through legislative records, media outlets, and social media. This process includes determining which major public claims (a) can be fact-checked and (b) ought to be fact-checked.
2. Find the facts by looking for the best evidence regarding the claim.
3. Correcting the record by evaluating the claim in light of the evidence, usually on a scale of truthfulness (Mantzaris 2018, 84).

In the digital age, fact-checkers often use various technologies, such as automatic speech recognition, news alerts, and translation tools, which typically depend on underlying AI technologies. Since fake news spreads six times faster than real news, these technologies could allow fact-checking organizations to provide more timely and extensive coverage than manual checks (Diakopoulos 2016; Owen 2017). But the authors notes other challenges, since automated claims verification could be AI's ultimate application to fact-checking, as many claims are not simply 'false' or 'true' but 'partially false' without additional context.

During the coronavirus pandemic, Full Fact created campaigns to analyze COVID-19 claims and automatically highlight errors, and developed industry-specific software for claim recognition, statistical analysis, and fact-checking based on pattern recognition and structured data (Sippitt 2019). DeFleur (2013) emphasizes the importance of understanding how fact-checking can be integrated with AI technologies to improve the public presentation of fact-checked results. Thus, this thesis will also explore how technology could enhance fact-checking through other AI applications.

Another point always considered within the problem of disinformation is bias, which is the tendency to favor one side of an issue or present data in an unfair or distorted way (Ahmad et al. 2023). Journalism distinguishes between information that claims to be impartial and content that explicitly expresses a point of view. Opinion columns, newspaper editorials, and political commentaries on television are designed to persuade the public to agree with the author's or commentator's perspective. For some authors, the position endowed by fact-checking content is akin to a quest for impartiality in information communication (Graves and Amazeen 2019).

Another issue in relation to bias lies in understanding the construction of disinformation, which refers to confirmation bias and motivated reasoning contents. Confirmation bias refers to individuals' natural, unconscious tendency to notice and accept information that agrees with or reinforces their existing beliefs while overlooking or questioning information that contradicts those beliefs (Roozenbeek and Van Der Linden 2021). For instance, sports fans may notice and remember close calls by referees that go against their team more than those that favor theirs.

Motivated reasoning, on the other hand, is the process of actively searching for and interpreting information in biased ways that reinforce current beliefs. This includes selectively looking for information that confirms pre-existing beliefs, failing to seek information that complicates or conflicts with these beliefs, cherry-picking evidence, and finding reasons to dismiss inconvenient facts and credible evidence (Roozenbeek & Van Der Linden, 2021). An example is sports fans inventing reasons to justify close calls that benefit their team while exaggerating the impact of calls that go against their team.

Part of the significant challenges in addressing the infodemic problem lies in the complexity of the disinformation environment, where journalism and fact-checking entities are only part of the solution. In health contexts, responses to disinformation and misinformation vary according to motivations. For instance, the decline in using traditional media for news can spread fake news while creating content promising artificial treatments for profit exemplifies current disinformation issues (Posetti and Bontcheva 2021).

In this context, Noortje Marres' work provides critical insights into fact-checking practices' limitations and potential pitfalls. Marres (2018) critiques the reliance on fact-checking language to combat opinion manipulation on the Internet and argues that they often enforce a problematic normative that aims to re-establish a hierarchy between knowledge and its presumed opposite, non-knowledge or anti-knowledge. Moreover, Marres highlights the risks of reinforcing societal dichotomies and amplifying public divisions between virtuous knowledge and sinful non-knowledge (Marres 2018, 432). Therefore, she calls for alternative, less judgmental ways of carving up public-issue spaces. She suggests that we test public

media for "experimental facts" whose truth value may fluctuate over time, fostering more inclusive and dynamic public debates.

In the realm of veridiction, author de Barros (2020) explains that the nature of relationships between existence and appearance determines various discourses: true, deceptive, secretive, or false. Within his enunciative framework, the recipient of the discourse selects a veridiction regime and shapes the audience's interpretation, leading them to believe or disbelieve in accordance with the agreed-upon verificatory contract. This interpretation relies on the audience's knowledge, beliefs, emotions, and the speaker's persuasiveness. When the interpretation primarily relies on the audience's beliefs and emotions, deceptive discourses are perceived as true, known as confirmation bias (de Barros, 2020).

To counter disinformation proactively, prebunking, a concept rooted in inoculation theory (Compton 2013), offers different avenues. Roozenbeek and Van Der Linden (2021) emphasize implementing strategies that tackle disinformation during its nascent stages. This approach involves developing psychological resistance against manipulative tactics akin to a "psychological vaccine" against disinformation. By preemptively exposing individuals to deceptive techniques, it is possible to equip them with tools to discern and resist manipulation.

As we summarize, the battle against misinformation on the internet is a complex one. While post-hoc attempts like fact-checking can address false claims, their effectiveness varies due to the complex factors influencing their efficacy. Therefore, effective communication strategies are essential for mitigating the impact of misinformation and promoting an informed online environment (Clayton et al. 2020; Pennycook, Cannon, and Rand 2018; Roets 2017; Thorson 2016; Vraga and Bode 2020). Although fact-checking has been primarily studied within political contexts, there is a growing need for research examining its impact on health (Sylvia Chou, Gaysynsky, and Cappella 2020).

In addressing the evolving media landscape, Singer (2023) discusses the changes in gatekeeping due to digital and social media, recognizing that information assessors and verifiers no longer control public domain content. The study recognizes that information assessors and verifiers no longer hold the same

level of control over public domain content due to the growing role of users and algorithms in shaping online activity:

The best that humans exercising journalistic oversight can offer is an antidote to misinformation: either through the further reporting and analysis that reporters provide, or through an exploration of the veracity of claims, and a debunking of those found to be untrue, that is the fact-checkers' specialty. This debunking is the result of what also can be seen as a gatekeeping function: the fact-checker's selection from among the myriad bits of nonsense ricocheting around the ether on any given day" (Singer 2023, 345).

Furthermore, Coddington, Molyneux, and Lawrence (2014), propose that fact-checkers adhere to a scientific notion of objectivity, testing hypotheses and arriving at evidence-based conclusions. Graves (2016) notes the global fact-checking movement, integrating journalistic, academic, and political-civic dimensions. Further comparative research is needed to understand this growing diversity, highlighting independence and transparency as fundamental attributes in the media landscape. As we transition to the next part of our literature review, we will delve into the generative media's contribution to sharing health fact-checking. This narrower focus aims to explore innovative ways to enhance the reach and impact of fact-checking in the digital age.

Section 2: Synthetic media contribution to share health fact-checking

In this second part, we examine the implications of new technologies to the previously discussed digital disinfodemic scenario. We will present the literature on using artificial intelligence (AI) in journalism and its associated ethical dilemmas and then address the potential contributions of these tools to the automation of health fact-checking. A broader understanding of AI-powered journalism addresses a specific aspect of the disinformation problem. Technology in this context pertains to practicality and instrumentalization and can be observed in how society organizes available pragmatic knowledge, inherent principles, and the context of use (Bayer and Osendorfer 2015). Simondon (1989) elucidates the complexities of this technological articulation, highlighting tools and machines marked by their purpose and contribution.

Several definitions of AI emerge from the specificities of various fields of knowledge. Stuart Russell (2022) categorizes eight definitions into two main groups: a) those referring to mental processes and reasoning and b) behaviors. In 1956, a workshop at Dartmouth College first introduced the term “artificial intelligence.” The workshop aimed to define the field as the science and engineering involved in creating intelligent machines, particularly those that can perform tasks requiring computer programs with human-like intelligence (McCarthy 2007). Due to the broad and imprecise nature of “intelligence,” David Geiger defines AI as the discipline focused on developing machines capable of performing tasks commonly associated with animal brains (Kaufman 2021).

One of the initial themes requiring further attention in this project is the concept of technologies and their role in automated fact-checking (Cushman et al. 2010). A barrier to automation is the lack of structured journalism in fact-checking, as many fact-checkers still rely on outdated blog platforms to publish their articles. However, the databases generated by hundreds of fact-checking initiatives provide access to valuable, structured data, which presents promising fields for automation, particularly in health information (Nakov et al. 2021; Santaella 2001).

While various natural language processing and machine learning techniques have been employed in Automated Fact-Checking (AFC), little attention has been

given to the quality of datasets used in these systems. Dierickx, Lindén, and Opdahl (2023) propose a practical framework for assessing and enhancing data quality in AFC systems to address this issue and align AI-based solutions with ethical journalism standards. By employing an interdisciplinary approach, the research identifies limitations related to ethical principles of truth, fairness, and transparency. The study emphasizes the need to ensure data sources' reliability and credibility while addressing challenges in maintaining data relevance over time. Furthermore, critically assessing data sources, including annotated data from Wikipedia and social media, is crucial in building trust between journalists and AI-driven systems. The following section will delve into the ethical issues surrounding artificial intelligence.

2.1 Artificial Intelligence Ethical Issues

Considering these challenging perspectives, our study seeks to identify ethically aligned approaches to technology, given the dichotomous directions this field can take, as discussed in the previous chapter. Internet governance presents numerous concerns and theoretical possibilities, with digital ethics being a critical focal point. This domain is also called information, computer, or data ethics (Floridi and Taddeo 2016). Business ethics literature offers sophisticated models to enhance our understanding of ethics, suggesting that decisions are connections between personality traits and situational characteristics (Trevino 1986).

Society is in constant flux, as are its challenges, including environmental and health inclusion issues. Digitalization is expected to address these challenges, but it also introduces new concerns (Floridi 2018). This section will discuss some fundamental aspects of these challenges. Ethical decision-making literature predominantly treats this category descriptively. According to Jones, unethical behavior is "illegal or morally unacceptable to the larger community" (T.M. Jones 1991, 367). It also involves understanding who holds power and data, who has expertise and knowledge, and who sets norms and decides their use, leading to ethical blindness:

Ethical blindness can be defined as the temporary inability of a decision-maker to see the ethical dimension of a decision at stake. The phenomenon can be

understood in three aspects. First, it builds on the assumption that people deviate from their values and principles. These values and principles are part of their identity, and they have tried to live up to them in the past. Second, ethical blindness is context-bound and thus a temporary state. It describes the psychological state of people with normal (or even high) levels of integrity and the ability for moral reasoning. But they cannot use these capacities when deciding for some reasons (often related to the situation, as we outline further below). Third, ethical blindness is unconscious. People who are ethically blind are not aware of the fact that they deviate from their values and/or that they cannot and do not access those values when making a decision (Palazzo, Krings, and Hoffrage 2012, 325).

The ethics of technology involves general principles and good practices to guide the AI ecosystem. As highlighted by Kaufman (2021), a significant challenge is the cultural variability in concepts such as justice and dignity. The Conference on Beneficial AI, held by the Future of Life Institute in 2017 and led by MIT cosmologist Max Tegmark, established 23 principles signed by 1,797 AI/Robotics researchers. These principles are divided into research, ethics and values, and long-term issues. In 2018, the AI4People event, Europe's pioneering global forum on AI's social impacts, was organized by the OECD in Brussels (Kaufman 2021). Bringing together over 50 experts, researchers, and representatives from industry and civil society, the concept of "beneficial AI" was outlined in the document "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." This framework addresses four fundamental points related to human dignity and flourishing: autonomous self-realization, human agency, individual and societal capabilities, and societal cohesion (Floridi et al. 2018).

The ethical use of automated journalism is a critical subject (Asaro and Wendell 2017; Crawford and Paglen 2021; Kelly 2010; Norvig 2002). The concern over biased or unfair machines is central to automated journalism (Søraa 2017). According to Royakkers et al. (2018), ethical and social issues vary by technology and relate to power balance, control (autonomy), human dignity, privacy, safety, distributive justice, and value collection, integration, analysis, and application. The balance of power concerns how digitalization influences market and government dynamics, with Big Tech often having a knowledge advantage over governments and crossing legal borders (T.M. Jones 1991). Society has become increasingly dependent on a few companies and their data-driven decisions.

Control (autonomy) is another critical issue, as the world is increasingly interconnected beyond geographical boundaries. Royakkers et al. (2018) argue that knowledge should be universally accessible, yet it remains geographically and culturally bounded. This highlights the need for ethical considerations in decision-making and the visibility of information. Control over technology also involves unpredictable effects, such as labor issues and technological consequences. Transparency in automated decision-making is crucial, but it presents a bias problem concerning who understands the technology (Royakkers et al., 2018). Transparency and insight into persuasive technology applications are essential for protecting autonomy (Royakkers et al., 2018, p. 132).

Privacy concerns include exchanging information without user control, pervasive monitoring, potential abuses of home digitalization, misuse of avatars, and unauthorized use of platform interactions (Royakkers et al., 2018, pp. 130-131). Privacy policies must ensure reliable algorithms that simulate "what if" scenarios to avoid ethical issues, considering the limitations on private space in the digital landscape (M. van Steenberg and van der Spoel 2021). Human dignity must be preserved in AI ethics, which relates to Kantian moral dilemmas, such as never treating humans merely as a means to an end. Digitalization should not dehumanize or cause mass unemployment. For example, in medicine, AI could lead to doctors becoming insensitive to death and detached from the consequences of errors:

Although robotics can provide great support in health care, entertainment, the police, and the army, if the technology is not applied within certain framework conditions, it can undermine human dignity. We are talking about the risk of objectification or instrumentalization of people, in other words, dehumanization (...) Critics fear that the danger lurking in creating more distance between an action and its consequences is that controllers make important, sometimes life or death decisions, as if they are playing a video game (Royakkers et al., 2018).

Safety concerns include digital crime, internet hacking, identity fraud, and psychological issues in virtual environments. An example is an avatar being indecently assaulted in the game Second Life (Royakkers et al., 2018, p. 134). Distributive justice involves issues such as profiling, presumption of innocence, exclusion, and checks & balances, which must be considered in digitalization. AI-enabled facial recognition can lead to wrongful accusations based on biased data

(Sutrop and Laas-Mikko 2012), reinforcing racial stigmas without the individual's knowledge.

Royakkers et al. (2018) also discuss AI in the human loop and the vertical versus horizontal perspectives on decision-making. Vertical decision-making involves top-down rules that can be automated, while horizontal decision-making requires collaborative deliberation, particularly in journalistic automation. This approach incorporates algorithms into human decision-making processes, considering the interplay of ethics, law, and sociology. Universal values should be anchored in law, such as human rights, but ethical norms and values, while part of the law, are not mandatory. The law often lags behind technological developments and may not always be just.

At this point, it is worth revisiting some previously discussed philosophical concepts. Hannah Arendt's approach highlights the importance of the public sphere, which has two dimensions: the space of appearances, where citizens act in concert through speech and persuasion, and the shared public world of human artefacts, institutions, and settings. Both dimensions are essential for citizenship, allowing individuals to act, reveal their identities, and establish reciprocal relationships (Arendt 2013). Arendt defines the public sphere with the following characteristics: 1. Artificial: a man-made construction; 2. A quality space where citizens can meet, exchange opinions, debate differences, and seek collective solutions to problems; 3. A distinction between public and private interests, as public interest, is not merely the sum of private interests.

Van der and van Steenbergen (2020) argue that the digital transition impacts Arendt's public sphere as technology becomes more than a tool. The shift in power relations through persuasion also leads to the loss of a shared world and increased personalization. They recommend understanding that digital innovation brings greater ethical responsibility for those involved in the innovation process, including privacy and inclusion norms. Some models are described (Table 2) with philosophical currents, where van Steenbergen presents examples to facilitate understanding these values that may also be implicit. Technological development can create or hinder values:

It not only considers the values of direct stakeholders but also of stakeholders who may indirectly be impacted by the innovation. For example, future generations or individuals cannot or will not use a service. The values of all stakeholders and potential tensions between them are iteratively investigated from a conceptual, empirical, and technical perspective (van der Stappen & van Steenbergen, 2020, p. 488).

Table 2. Recommendations for Dealing with moral Dilemmas

	Utilitarianism (Bentham/Mill)	Deontology (Kant)	Virtue (Aristotle)
Motivation	The aim is to produce the best.	The aim is to perform the right action.	The aim is to develop one's character.
Definition of Ethical Conduct	Ethical conduct is the action that will achieve the best consequences.	Ethical conduct involves always doing the right thing; never failing to do one's duty.	Ethical conduct is whatever an entirely virtuous person would do in the circumstances.
Deliberative process	What kind of outcomes should I produce (or try to create)?	What are my obligations in this situation, and what are the things I should never do?	What kind of person should I be (or try to be), and what will my actions show about my character.?

Value Sensitive Design (VSD) does not define itself in any classical theory, but we can assume it is closely aligned with Virtue Ethics. VSD emphasizes integrating human values into technological design, focusing on developing good character traits and virtues, which is consistent with the goals of Virtue Ethics. The deliberative process in VSD involves stakeholders understanding and incorporating their values, similar to how Virtue Ethics considers what a virtuous person would do in a given situation. VSD asks questions like "What kind of values should this technology embody?" and "How can we design this technology to promote human flourishing?" which mirrors the process of considering actions taken by a person with a virtuous character. While VSD incorporates aspects of Utilitarianism and Deontology, its emphasis on values and character development aligns it most closely with Virtue Ethics, fostering an ethical and virtuous society.

2.1.1 Value-Sensitive Design Approach

With the adoption of ethics as the basis for technological innovation, some approaches have gained prominence. This is the case of the need for IT to embrace

usability and aesthetics. The Human-Computer Interaction - HCI field gained notoriety, as well as participatory design methods. Participatory Design was developed in Scandinavia in the 70s, proposing low-fidelity prototypes and commitments to democracy and equalization. This proposal was initially more connected within organizations. The Value-sensitive design got some basic understanding from them but did apply it to other scenarios, such as home, public life, school, etc. (Borning, Friedman, and Kahn 2004).

In this context, we will approach computing ethics through value-sensitive design, which we will address in this chapter. Our perspective is closer to the scholarship on Human-Machine Communication (HMC), defined as “an emerging area of communication research focused on the study of the creation of meaning among humans and machines” (Schilling et al. 2020, 71). This perspective proposes to analyze the interactions between professionals and technical artifacts as “communicative subjects,” considering the implications for programmers, journalists, and society. Diakopoulos (2019) also presents it from a possibility of automated journalism, where human tasks are combined with algorithms that adapt to the potentiality of technology and shelter their limitations.

In journalism, ethical standards are rooted in the notion of social responsibility. This encompasses news content, the role of news media in society, and the responsibility news media holds towards the public (Bardoel and dHaenens 2004). While ethical journalism is primarily a matter of practice, it is guided by universally acknowledged principles, including the importance of truthfulness. This means presenting verified facts based on reliable sources, reporting accurately, and providing well-balanced information with fairness, independence, and non-partisanship (Hafez 2002). Another critical principle in journalism is objectivity, an essential aspect of professional self-perception and identity (Deuze 2005). However, this concept often faces criticism as it is viewed as an ideal or even a myth due to its dependence on the subjective nature of individual journalists (Ward 2019). The choices made by journalists, such as selecting a topic, an angle, sources, and the narrative, further highlight the impossibility of complete objectivity, as they involve human and organizational decision-making (Tong and Zuo 2021).

Another way to understand the responsive use of AI is through the area of policies for technological innovations. According to Borrás and Edquist (2013) innovation policy instruments used by governments and public agencies must be carefully designed and combined into "policy mixes" that address specific problems within the innovation system. These instruments can be categorized into regulatory, economic, financial, and soft instruments. The choice and design of these instruments are influenced by political processes and require a deep understanding of the innovation system's needs. The systemic nature of innovation policy instruments and the importance of customizing them to the unique problems of each innovation system are emphasized.

Additionally, Edler and Fagerberg (2017) describe innovation policy instruments as encompassing a diverse range of strategies designed to enhance research and development (R&D) and stimulate innovation within firms. These strategies include fiscal incentives for R&D, direct support for firms' R&D and innovation activities, and policies to improve training and skills. Furthermore, entrepreneurship policy and technical services and advice play significant roles. Cluster policy and policies to support collaboration and innovation networks are crucial for fostering a conducive environment for innovation (Laatsit, Grillitsch, and Fünfschilling 2022).

Among the various methods to align AI development with ethical principles, the "AI for Social Good" (Umbrello 2021) stands out as a notable approach. Detailed by Floridi, Cowls, King, et al. (2021), this method identifies seven crucial elements for designing AI systems with societal benefits in mind. These elements include testing and phased implementation, safeguarding against the distortion of predictive models, delivering context-specific interventions, providing context-aware explanations with clear objectives, ensuring privacy and securing data subject consent, maintaining contextual fairness, and facilitating user-friendly interpretations (Floridi, Cowls, Beltrametti, et al. 2021, 1773). Though these elements are distinct, they are interconnected and not arranged in a hierarchical order. They align with the EU High-Level Expert Group on AI's ethical principles—respect for human autonomy, prevention of harm, fairness, and explicability. By detailing these seven elements, the AI4SG approach seeks to translate broad ethical

values into specific norms and design guidelines, emphasizing the integration of ethical considerations into AI development to foster technological innovation that also promotes social welfare.

Understanding that there is a tremendous challenge in seeking to make a difference, leaders propose innovation models that include values and how to transfer them into solutions that include ethical decisions as a phase to be exercised (Spiekermann 2015). The adoption of value models has been a way to incorporate ethics into technological innovation, some of which are known as the value survey model (McCoy 1969), the value pie model (R. Pereira and Baranauskas 2014), the theory of fundamental values (Schwartz 2012) and VSD - Value Sensitive Design (Friedman and Hendry 2019). The model we believe to be most appropriate for our study is the VSD, which is defined as “a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process” (Borning, Friedman, and Kahn 2004, 2).

Value Sensitive Design (VSD), developed by Batya Friedman and Peter Kahn at the University of Washington in the late 1980s, originated from the fields of human-computer interaction (HCI) and information systems design. It represents a principled approach aimed at integrating human values into technology design, primarily through engaging with stakeholders (Friedman, 1996). VSD advocates for a philosophical perspective on the interaction and relationship between humans and technology, challenging the views that technology is either deterministic or value-neutral. Instead, it posits that technology inherently carries values. This recognition has led to the development of various stakeholder theories and methods focused on creating technologies that yield positive impacts (Friedman 1997).

Bestowing to the model, VSD takes an interactive stance about technology and values in terms that people shape technology, and technology shapes people. The view is that technology is not good, bad, or neutral but reflects the values of designers and implementors, intended or unintended. VSD takes values into account in the design process in these concepts: a) define context, b) clarify project values, c) identify stakeholders, d) identify benefits & harms, e) identify and elicit values, f) operationalize values, g) investigate value tensions, f) formulate design

and implementation requirements (Borning, Friedman, and Kahn 2004). This model will be better clarified to be used in our case analysis.

In the sequence that innovation projects have a digital core, the ethical responsibility of those involved includes raising awareness of social interest aspects in this process, considering that implementing a technological artifact can enable and hinder fundamental values. To help in this, VSD is "a theoretically grounded approach to technology design that accounts for human values in a comprehensive and principled way throughout the design process" (Nakamoto 2009, 349; Friedman and Hendry 2019). Thus, its instrumental aspects include functionality, reliability, and ease of use, but are mandatorily allied with the moral values of individuals and societies (Flanagan, Howe, and Nissenbaum 2008).

To understand the application, it is worth citing the meanings of value, which, for Friedman and Hendry (2019) represents "what is important to people in their lives, with an emphasis on ethics and morality" (2019, 4). For Rokeach (1973), it is an enduring belief that conduct or existence is personally or socially preferable to an opposite mode of behavior or state. Schwartz (2012), meanwhile, indicates that values are beliefs, refer to desirable goals that motivate action, transcend specific activities and situations, serve as standards or criteria, are ordered by relative importance, and guide steps. Other authors (M.T. van Steenbergen, Cathelijn 2022) further categorizes types of values along a dichotomous line: a) economic/monetary versus psychological, b) intrinsic versus Instrumental, c) personal versus organization /public, and d) moral versus non-moral.

VSD integrates the values to consider not only direct stakeholders but also those who may be indirectly affected by the proposed technology. For example, future generations or individuals cannot or do not want to use a service (van der Stappen and van Steenbergen 2020). The values of all stakeholders and potential tensions between them are "iteratively investigated from a conceptual, empirical, and technical perspective" (van der Stappen and van Steenbergen 2020, 488). Borning, Friedman, and Kahn (2004) suggest that including direct and indirect stakeholders from the start is a valuable contribution of VSD. However, a VSD problem is that the number and different kinds of stakeholders can be inserted without limits and in numbers that can't be controlled (Borning, Friedman, and Kahn 2004).

M. van Steenbergen and van der Spoel (2021) present the model in three levels: conceptual, empirical, and practical. At the conceptual level, VSD considers including relevant stakeholders and values based on existing literature and knowledge. At the observed level, various stakeholders' perception of these values is studied through interviews, focus groups, or tests, leading to the development of norms. In practical use, the norms are translated into technical design and prototyping. VSD has a tripartite approach, which includes a first edge as the investigation of concepts, which involves identifying values and relevant stakeholders using philosophical methods. The second edge is represented by empirical research, which presents norms employing applied social sciences. The third edge brings the technical investigation requirements, with technological methods and infrastructure being used. M.T. van Steenbergen, Cathelijm (2022) cites an example (Figure 4) where a welfare value is represented in norms and requirements to apply a value-sensitive design.

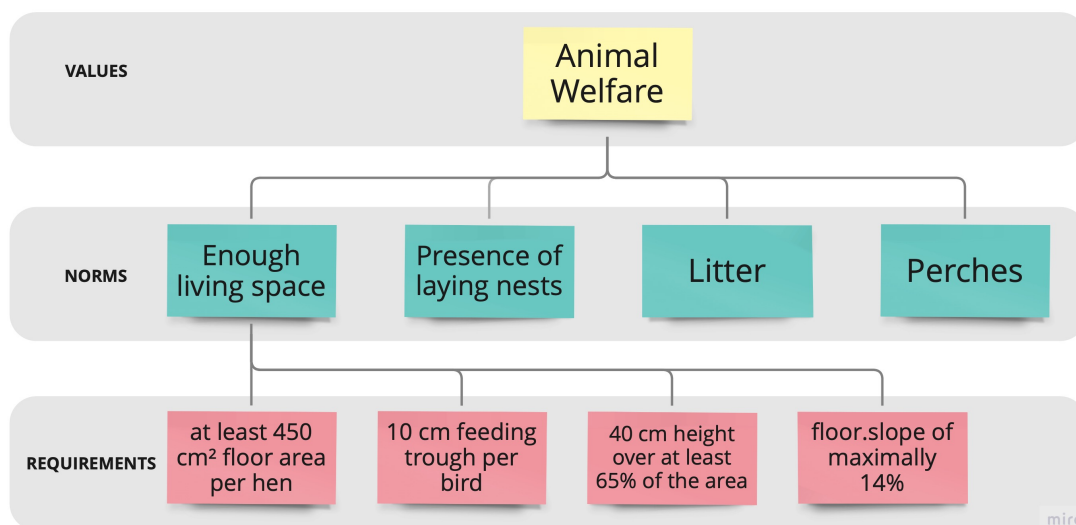


Figure 4. Example from value to requirements (M.T. van Steenbergen, Cathelijm 2022)

Understanding the challenge of collecting relevant principles, some authors establish criticisms of these models that do not have an easy solution. Millestadt (2019), for example, makes an exciting approach that values alone could not guarantee ethics in AI for medicine. According to the author, in medicine, the development of "AI lacks (1) common goals and fiduciary duties, (2) professional

history and standards, (3) proven methods for translating principles into practice, and (4) robust mechanisms for legal and professional accountability (Mittelstadt 2019).” There are some weaknesses of the approach to ethics that can be transcribed to AI regarding a relative lack of professional history and well-defined norms of 'good' behavior for journalism automation. Stronger professional norms for developing this area go through better knowledge of AI by journalism professionals themselves, but unfortunately, this is not a simple task.

Nevertheless, systems are often created by multidisciplinary and multinational groups. The great diversity of interests and people involved in searches for shared values and norms go to a shallow level (Dourish and Bell 2011). The promise of AI of possible job substitution also makes multidisciplinary difficult, which is understandable for a work environment with very fragile securities. This nebulosity makes AI entangled in the ethical and political dimensions of the vocations and practices in which it is embedded. As Mittelstadt (2019) highlights, “AI ethics is effectively a microcosm of the political and ethical challenges faced in society (Mittelstadt 2019, 10).” For the next chapter, we will try to settle a bit of the challenge of transparency in explaining concepts of AI and technology so that lay readers from this thesis understand the basic concepts embedded in this great dilemma.

VSD's approach of not endorsing any specific ethical framework while recognizing universal values leaves room for the acceptance of controversial moral ideologies, such as Nazism (Albrechtslund, 2007). The approach's creators note that while some ethical frameworks (like utilitarianism or deontological ethics) may support the realization of certain values, others (such as those based on virtue ethics) might not (Friedman and Kahn Jr., 2003). This issue continues to spark discussions and disagreements within the VSD community. As explored in the previous part of this thesis, the contention stems from VSD's reliance on somewhat simplistic ethical theories, including deontology, virtue ethics, and utilitarianism.

Value Sensitive Design (VSD) programs are guided by eight key considerations outlined by Friedman et al. (2008), designed to implement its iterative methodology effectively:

a) Starting Point: VSD begins with a focus on either a value, technology, or the context of use, depending on which aspect is most pivotal to the designer's objectives or interests.

b) Stakeholder Identification: It's crucial to systematically recognize both direct stakeholders, who interact with the technology or its output, and indirect stakeholders, who are impacted by the technology without direct interaction.

c) Assessing Harms and Benefits: For each stakeholder group, identify how they could be positively or negatively affected by the technology.

d) Value Mapping: Harms and benefits should be mapped to corresponding values, acknowledging that this process can reveal complex relationships between harms, benefits, and multiple values.

e) Conceptual Value Investigation: This involves developing precise definitions for key values, drawing from philosophical literature to refine these concepts and explore how they can be converted into norms and, subsequently, design requirements.

f) Value Conflicts: Designers should identify potential value conflicts, not as binary oppositions but as constraints within the design space, recognizing common conflicts such as accountability versus privacy or environmental sustainability versus economic development.

g) Technical Investigation of Value Conflicts: Design trade-offs often involve adjudicating between multiple or conflicting values. Designers are tasked with clarifying how these trade-offs reflect underlying value conflicts and their impact on different stakeholder groups.

h) Flexibility for Unanticipated Consequences: To manage unforeseen consequences and value conflicts, designs should incorporate flexibility within the technical architecture to allow post-deployment modifications.

These considerations are not meant to serve as a rigid sequence of steps but as a flexible framework to support the integration of VSD into engineering and design processes. This approach is intended to complement, not replace, existing methodologies, encouraging adoption and adaptation within engineering and design cultures, as exemplified by initiatives like the IEEE's Ethically Aligned Design standards.

2.1.2 Technological ethical perspectives and challenges

Among the moral challenges the Internet presents, the use of data and the ethics surrounding these usage models remain central themes. One pertinent concept is "surveillance capitalism," developed by (Zuboff 2019), which addresses the quest of powerful corporations to predict and control behavior. In her pioneering work, Zuboff explores the phenomenon where digital architecture serves as a global mechanism for behavior modification, commodifying human actions. She defines this ubiquitous computational architecture as "in which automated machine processes not only know our behavior but also shape it at scale" (Zuboff, 2019, p. 8). This critique underscores how trust, autonomy, and identity are treated as tradable commodities.

In moral philosophy, ethicists typically assume an ethical standpoint from which decisions can be evaluated, which is also relevant to technology developers. Palazzo, Krings, and Hoffrage (2012) Acknowledge that moral viewpoints vary based on philosophical underpinnings—such as Kantian duty versus utilitarian calculus. However, they concur that people can adopt an objective, unbiased criterion for weighing arguments and decisions (Hunt and Vitell 1986). Tenbrunsel and Messick (2004) note that ethical considerations can fade in decision-making, often due to factors like euphemisms (Chugh, Bazerman, and Banaji 2005). From the perspective of machine cognition, moral judgments are emotional, whereas human judgments are culturally learned, as our brains are not blank slates (Pinker 2003).

Weick (2005) posits that sensemaking builds on previous experiences that shape our worldview. Palazzo, Krings, and Hoffrage (2012) present a model of the individual sensemaking process dominated by specific frames and contextual pressures—organizational, situational, and institutional (Palazzo, Krings, & Hoffrage, 2012, p. 328). This process highlights how external pressures influence decision-making and the articulation of values.

In technological development, journalists' understanding—or lack thereof—of technology can influence sensemaking and introduce biases. The interplay between

technology adoption and journalism reveals challenges, including a misalignment between journalists and the exact sciences and barriers posed by language and moral purposes. Despite its benefits, Y. Yao et al. (2017) points out that AI in journalism has dangers, such as technologies that generate opinions mimicking human comments, potentially promoting misleading information. Morozov (2008) emphasizes that immense wealth in the tech sector is concentrated among a few investors, which does not necessarily benefit global economic growth but rather establishes data and AI services as gatekeepers:

In the economic field, we see immense wealth being accumulated by just a handful of investors who were smart and quick enough to invest in the sector. The mechanics of this industry, however, are not necessarily conducive to the recovery of global economic growth—the concentration of data and, by extension, AI services in the hands of just a few companies may cause them to become gatekeepers (and potentially profit-seeking gatekeepers) of the new digital economy (Morozov 2018).

Distinguishing between transactional and use value is crucial in various contexts. Transactional value focuses on monetary aspects, while use value assesses an item or service's adequacy in meeting specific needs. Jensen et al. (2012) comprehensively examine value dimensions, including customer, economic, social, environmental, relationship, and surplus values. The notion of surplus value is particularly relevant in economic contexts, highlighting the profit derived from labor beyond its direct cost (Acemoglu and Restrepo 2018).

Key concepts in this area include “disruptive innovation” (Christensen 1997) and the “anxiety of automation” (Akst 2013). The fear of technology replacing jobs is often exaggerated, especially in sectors where human oversight remains crucial. There is no consensus on technology's impact on labor, with some predicting greater inequality and task substitution rather than job elimination (Frey and Osborne 2013; Arntz, Gregory, and Zierahn 2016). Hidalgo et al. (2020) explore how people perceive technology's impact on jobs compared to other forces, noting a preference for displacement attributed to technology over foreign workers. They emphasize that technology will undeniably affect the future of work alongside global value chains, economic concentration, education levels, and international migration:

The fact that technology will affect the future of work is undeniable. But technology is not the only force affecting labor. The future of work also depends on global value chains (Gereffi, Humphrey, and Sturgeon 2005), the increasing concentration of complex economic activities (Balland et al. 2020), the rising education levels of the Global South (Roser and Ortiz-Ospina 2016), and international migration (Rapoport 2016). To better manage this impact, we need to understand how people react to the impact of technology on jobs compared to other forces (Hidalgo et al. 2020, 10).

The integration of AI in journalism, exemplified by the Associated Press's adoption of AI roles, demonstrates potential benefits such as reducing workload while maintaining journalistic integrity (Barbizan 2021). Domingos (2015) cautions against overestimating AI, noting limitations in machine learning and the importance of data management and technical capabilities. Pearl and Mackenzie (2018) argue for the need to make machines understand causality, as traditional models fail to capture cause and effect well, which is essential for comprehensive data analysis. As Pearl clarifies:

Most relevant for us is the question of what comes next. How do we extract meaning from all these numbers, bits, and pixels? The data may be immersed, but the questions we ask are simple. Is there a gene that causes lung cancer? What kind of solar system is likely to harbor Earth-like planets? What factors are causing the population of our favorite fish to decrease, and what can we do about it? In certain circles, there is an almost religious faith that we can find the answers to these questions in the data itself if only we are sufficiently clever at data mining. (...) The questions I have just asked are all causal, and causal questions can never be answered from data alone. They require us to formulate a model or process that generates the data or some aspects of that process. Anytime you see a paper or a study that analyses the data in a model-free way, you can be certain that the output of the study will merely summarize and perhaps transform but not interpret the data (Pearl and Mackenzie 2018, 351).

Broussard (2023) highlights AI's inherent biases and the necessity for collaborative algorithmic accountability reporting to mitigate these issues. She underscores the importance of transparency and rigorous testing for accessibility, advocating for responsible and transparent use of AI in journalism. The Markup exemplifies data-driven investigations that promote transparency and accountability in journalism (Yin, 2022). As Broussard notes:

However, it is important to remember that just because we use the term “black box,” it’s not that it’s impossible to describe what happens inside the model. The ability to explain it is limited by the descriptive abilities of the people in the conversation, the context of the conversation, and the mathematical background of the people in the conversation. This stuff is hard to understand, and it takes

focus and a little bit of work. It is well within the reach of everyone, but I won't lie and pretend it doesn't take effort. Let's dig in (Broussard 2023, 10).

In AI governance, scholars emphasize the need for regulatory frameworks to address issues like mass surveillance and data colonization (Cheney-Lippold 2017; Taeihagh 2021). The ethical use of AI in journalism involves balancing competing values and ensuring responsible organizational processes (Díaz 2022; Helberger et al. 2022). Recent advancements, such as the draft AI Act ("European Commission" 2021), highlight the importance of clear usage policies and accountability.

Floridi et al. (2018) propose a unified ethical framework for AI, incorporating principles from classical biomedical ethics and adding explicability as a new enabling principle. This framework has significantly influenced AI ethics research and guidelines, such as the European Commission's 'Ethics Guidelines for Trustworthy AI': "Ethical principles in AI can be synthesized into only five, with four coming from classical biomedical principles—beneficence, non-maleficence, autonomy, and justice—added to a fifth principle, explicability, exhibited as a new enabling principle for AI" (Floridi & Cowls, 2019, p. 8). The next section of this thesis will delve into the ethical challenges posed by deepfake technology, exploring its implications for digital ethics.

2.1.3 Deepfake as altercate to digital ethics

Within the realm of ethical challenges, it is essential to note the characteristics of deep-fake technology that facilitate real-time and widespread dissemination in digital channels. While deep-fake technology offers certain advancements, it poses significant dangers, particularly within the field of disinformation. Deep-fake videos replace the face and speech of real individuals with manipulated content. As Yao et al. (2017) detail, although these procedures are somewhat labor-intensive, they are not impossible, and can be executed using powerful computers, publicly available applications from Google Corporation's Artificial Intelligence division, original videos, and image manipulation techniques. With the evolution of technology via neural networks to identify voice and motion patterns, such fake content is becoming increasingly convincing in deceiving people (Y. Yao et al.

2017). This content proliferates across platforms such as Facebook, Twitter, blogs, websites, and various types of text, audio, and video content.

The decay of factual truth, exacerbated by the use of this technology, interacts toxically with societal prejudices (Chesney and Citron 2019), and deepfakes significantly intensify this problem. The construction of deepfake videos (Chesney and Citron 2019; Gibney 2017), which can attribute false statements or actions to individuals, raises serious ethical concerns about manipulation and persuasion. Companies are increasingly racing to find solutions to both create and detect synthetic media (Gibney 2017; Marra et al. 2018; Tariq et al. 2018).

Deepfake technology has already permeated daily life, enabling the manipulation of videos using devices such as mobile cameras. This advancement allows for the seamless superimposition of one individual's deepfake onto another, creating the illusion that they are saying or doing something they are not. The availability of open-source deepfake tools, such as DeepFaceLab on GitHub, has contributed to the proliferation of deepfakes (K. Liu et al. 2020). Despite their potential, significant regulatory concerns exist, particularly regarding their misuse and manipulation. For instance, by intentionally reducing the video quality of deepfake content shared on platforms like WhatsApp, it becomes more plausible for viewers to believe the manipulated video due to lowered visual fidelity, exploiting perception limitations (Diakopoulos and Johnson 2021).

Public attention to deepfakes surged after a Reddit user developed a tool to insert celebrity faces into pornographic scenes (Ellis 2018). Another pioneering app, FakeApp, provided a desktop service for creating deepfakes using deep learning (Güera and Delp 2018). A New York Times reporter created a semi-realistic deepfake video of his face on actor Chris Pratt's body using 1,861 images of himself and 1,023 of Pratt (Roose 2018). Misusing such technology is dangerous, as negative, and novel information tends to capture public attention. Citizens are attentive to threats and negative messages (R. Meyer 2018). Studies indicate that hoaxes spread rapidly, facilitated by filter bubbles reinforcing existing beliefs:

(...) filter bubbles further aggravate the spread of false information. Even without technology, we naturally surround ourselves with information confirming our beliefs. Social media platforms supercharge this tendency by empowering users to endorse and re-share content. Platforms' algorithms highlight popular

information, especially if it has been shared by friends, and surround us with content from relatively homogenous groups (Chesney and Citron 2019, 1768).

Chesney and Citron (2019) outline key ethical concerns regarding the misuse of deepfakes, such as false videos depicting public officials taking bribes, insulting politicians, or fabricating statements on controversial issues (p. 1776). An example includes a famous lip-syncing deepfake of Obama (Figure 5), altered by Jordan Peele to use profanity (Kietzmann et al. 2020). Another notable instance involved a manipulated video of President Kim Jong Un uploaded to YouTube, highlighting the dangers and implications of this technology (Broinowski 2022).



Figure 5. Lip-syncing example from a famous internet Deepfake video

Deepfakes also pose risks to news organizations, necessitating regulation and reliable authentication methods to protect the integrity of media (Chesney and Citron 2019, 1776). This issue intersects with broader discussions on human and post-humanistic design, particularly concerning virtual avatars (Deleuze, 1992; Santaella, 2010; Søraa, 2017; Tlostanova, 2017).

Audio deepfakes and voice-swapping are also significant concerns. For example, fraudsters mimicked a CEO's voice to trick a manager into transferring \$243,000 (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017). This technology allows for age and accent modifications and has been used in multilingual advertisements (Almutairi and Elgibreen 2022).

A manipulated video of Nancy Pelosi (Figure 6), appearing intoxicated, was widely circulated, highlighting the potential for deepfakes to damage reputations and influence public opinion (Buo 2020). This underscores the need for vigilance in verifying media authenticity and the consequences of deepfake misuse in political discourse (Ivancevich 2022).



Figure 6. Nancy Pelosi manipulated to appear intoxicated

Since 2017, China's Xinhua news agency has employed AI journalistic anchors with facial movements and translation capabilities (Xuya Wang and Zhu 2022). This trend has spread to broadcasters in Russia, the UAE, Brazil, and Thailand. Reuters partnered with Synthesia in 2020 to create an AI-generated presenter (Figure 7) for real-time sports bulletins during soccer games (Chandler 2020). Government initiatives in this area are less common than private ones, though the potential for deepfakes in governmental control warrants further investigation (Turek 2019).



Figure 7. Reuters Virtual Reporter Prototype

Kazakhstan's Atameken Business news channel introduced an AI news anchor (Figure 8), Ali Sanj, modeled on actor Sanjar Madi (Ashimova et al. 2022). This development coincides with a noteworthy event in the country, where the Kazakh government recently declined to accredit two journalists affiliated with an alternative radio service. This juxtaposition raises questions about the role of AI in journalism and its potential implications for media freedom and diversity in Kazakhstan.



Figure 8. AI Kazakh's Atameken Business News Anchor

The rise of deepfake technology presents significant challenges to digital ethics. Its potential for the widespread dissemination of manipulated content threatens

factual truth and exacerbates existing prejudices. Misuse of deep learning, whether in fake videos or manipulated voices, can harm individuals, manipulate public opinion, and deceive audiences. Efforts to regulate, detect, and authenticate deepfake content are necessary to preserve information integrity and protect society from malicious exploitation. On the other hand, the exploration of virtual characters and AI-generated news anchors further blurs the lines between reality and fabrication, necessitating better examination and ethical considerations. The subsequent chapter will delve into these issues in greater detail.

2.1.4 Uncanny Valley: how to present a virtual character?

In the realm of avatars, technology offers various avenues for enhancement, including integrating computer graphics and 3D elements. This cutting-edge 3D technology employs different methods, such as scanning, to capture real-world objects and transform them into three-dimensional characters (Grigorovici and Constantin 2004). These characters are constructed using a mesh structure, wherein multiple points represent distinct parts of the object. By connecting these points, a mesh is formed. Subsequently, this mesh can be optimized to remain lightweight enough for real-time usage. It finds extensive application in gaming and animation, enabling the creation of lifelike characters. Notably, Meta Unreal is an innovative technology that is democratizing the process of producing more realistic characters (D. Sharma and Sharma 2023).

Another technique employed in crafting 3D characters is photogrammetry, which involves capturing the physical attributes of real individuals and converting them into three-dimensional representations. The intricate process of modeling and texturing is necessary to replicate an individual's likeness in the virtual space accurately (Akaslan and Ulukavak 2022). Surprisingly, even mobile devices can be utilized to create and digitize a person's likeness for integration into Meta. Furthermore, the movements of the 3D characters can be animated using performance data obtained from sources such as Kinect on mobile phones (Latoschik et al. 2017). Skilled 3D animators possess a deep understanding of anatomical concepts, enabling them to accurately manipulate and deform the

mouth, face, and other facial features to simulate specific phonemes and expressions.

The creation of characters in media allows for the establishment of emotional connections. Individuals can shape these characters to align with their desired portrayals when they are customizable. The rise of virtual influencers provides some evident examples: Lu (Figure 9), a digital personality brought to life by Magazine Luiza, a prominent Brazilian retail company—is regarded as the leading virtual influencer globally, boasting an impressive fan base of 31.2 million followers across various social media platforms (Hiort 2022).

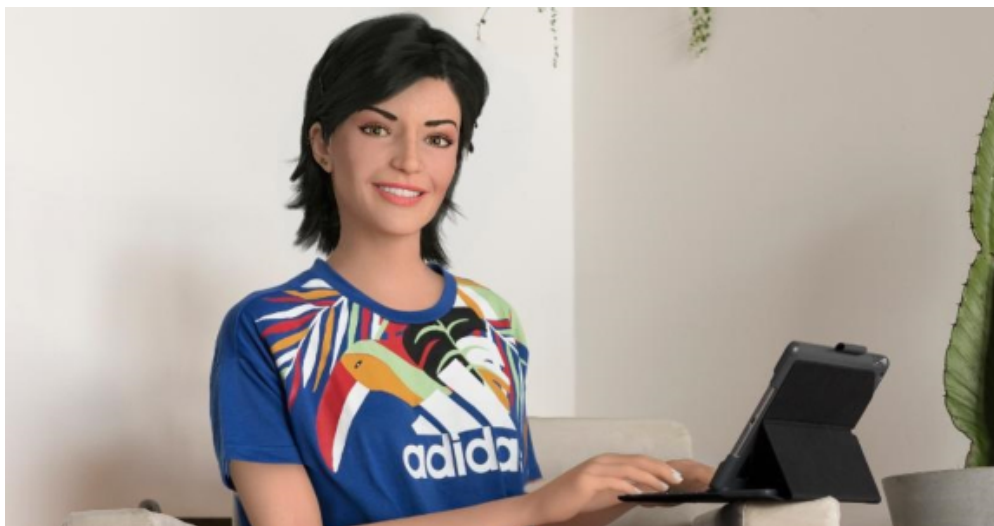


Figure 9. Lu for an advertising campaign. Image credit: Disclosure/Magalu

According to HypeAuditor (2022), virtual influencers possess three times the advertising impact in videos compared to real influencers, yet the exact reasons for this phenomenon remain undefined (da Silva Oliveira and Chimenti 2021). One possible explanation could be the novelty of using virtual characters, as their production has become more accessible and efficient. The advantages of utilizing 3D technology over deepfake techniques lie in its more significant level of customization.

In their article titled "The Uncanny Valley," Mori, MacDorman, and Kageki (2012) they discuss the concept of the uncanny valley in robotics and human perception. The uncanny valley refers to the phenomenon where a robot or artificial character's likeness to a human becomes increasingly realistic but reaches a point

where slight deviations from human behavior or appearance elicit a strong sense of unease or repulsion in observers. A robot lacking human-like features is generally accepted without causing discomfort. However, as more human elements are incorporated, such as skin, hair, and a realistic appearance, people's expectations shift towards anticipating human-like behaviors. Any deviations, such as strange mouth movements or faulty motions, become more noticeable, and the character is no longer perceived as human. Consequently, the affective response towards the character diminishes, leading to disliking or repulsion.

The challenge lies in striking the right balance between realism and avoiding the uncanny valley. The more realistic a character appears, the more important it becomes to eliminate any subtle irregularities that might trigger negative reactions. It is easier to create cartoons or animations, as the realism required for human-like characters poses significant difficulties. Achieving convincing human qualities in 3D characters remains a complex task. This is why deepfake technology is often employed to deceive individuals in news stories, as it relies on manipulating videos of real people within a network (Mori, MacDorman, and Kageki 2012). Understanding the uncanny valley concept is crucial for creating visually appealing and relatable artificial characters.

By carefully navigating the challenges associated with realism, creators can strive to minimize the sense of strangeness and enhance the acceptance and positive emotional response towards these characters. This understanding sets the stage for a deeper exploration into automation, generative media, and AI advancements.

2.2 Automation, Generative Media, and AI

In this step, we will understand some of the more technical concepts of a technological proposal for the challenge of health disinformation. Within what has been presented so far, it was possible to understand that there are different perspectives on understanding ethics in technological innovation. The technical perspective looks for the ethical consequences of algorithms on technical methods

and how systems are developed, focusing on improving the practices of building, modifying, or governing algorithmic systems (Ehsan et al. 2022). References to this interference are given in technical documentation and organizational accountability practices (Mitchell et al. 2019), audits of algorithmic systems (Buolamwini and Gebru 2018), demonstrations of the dangers of big data (Koch et al. 2021), and algorithmic justice (Smith 2022).

The contextual perspective seeks relations of ethics in the social and political relations with which the system interacts, acting on organizational roles and values in AI ethics (Amrute 2020), redistribution of social power (Raghavan et al. 2020), studies in social and political biases (Noble 2018), and recapitulating a worldview (Birhane 2021). The infrastructural perspective examines the social circumstances that make technology possible in the first place. The core lies in the unequal relationship of algorithmic systems (Leonardi 2013) to the existing socio-technical arrangements of classes and genders that constitute the database (Mitchell et al. 2019).

Further discussion is achieved by presenting specific practical applications and the breakdown of empirical work through a qualitative analysis of the phenomena and attributing their meanings. The concern relies on the necessity that technology's collective and individual impacts on media practitioners be better documented and that well-established literacy about automation could contribute to multidisciplinary automated journalism practices. This approach should not be considered a technocentric (Peeters et al. 2021) and simplified view that involves invisible journalist actors and prioritizes the technical work. Still, to not consider technology a great salvation and ignore its problems, our approach will bring the theories that analyze the ethical use of AI and applied models for the core values that should be embedded in journalism.

Nils Nilsson defines that Artificial Intelligence applications encompass some main objectives of reproducing human capacities, such as search, planning, knowledge representation, perception, ability to move objects (robotics), natural language programming (NLP), and learning (Bender 2013; Martinez-Plumed et al. 2018; N.J. Nilsson 1998). Applications also usually correspond to various subfields within these objectives (N. Nilsson 2014). When we talk about synthetic media, we

are faced with understanding the intuitive way in which robots generate dialogues with humans and if it best represents that purpose. Despite some challenges, the world has accompanied exciting developments in Natural Language Processing (NLP) and Computer Visions in recent years. To this end, researchers on both fronts - Linguistics and Computer Science - rely on the foundations of various disciplines, such as the Philosophy of Language, Social Communication, Cognitive Sciences, Psychology, Logic, and Mathematics, to develop computer systems with automated communication (Folger and Stein 2017).

Various perspectives of intelligence - perception, thinking, learning, invention, etc. - can be understood by their contribution to our capacity for efficiency. Since the beginning of IA, intelligence in machines has been defined by the extent to which their actions can achieve their goals (Stuart Russell 2021). In this perspective, Wiener had already evaluated that the machine had outperformed its creator in Arthur Samuel's chess game and that moral and technical consequences of automation should be considered (Wiener 1960, 1355). In short, the author pointed out that if we use, to achieve our goal, a mechanical agent in whose operation we cannot effectively interfere. it is better to be sure that the goal instilled in the machine is our desired goal.

When we delve into the relevance of technology and its dichotomies, it becomes necessary to unravel some essential concepts about significant categories of algorithms within machine learning. This is a subfield of Artificial Intelligence whose goal lies in the development of techniques that enable computers to learn rules and patterns from large data sets, or big data, as they are called. For this chapter, the technology will be analyzed precisely within the models applied to communication, with increasing importance to learning algorithms. As we have seen, algorithms have linear inputs and output

Machine learning subverts this input-output logic. In learners, data and the output generate algorithms to learn how that output was found. With machine learning, then, computers write their own algorithms, thus preventing humans from the process of programming them. When it comes to machine learning, the more data we have, the more it is possible to learn effectively. Paulo (Domingos 2015) states: "(...) the more data we have, the more we learn. If there is no data? There is

nothing to learn. Many data? A lot to learn. Therefore, machine learning is popping up everywhere, driven by the complexity of data(Domingos 2015)."

While we can point out the internet and technology as a symbol of democracy, we cannot neglect that they are linked to the new giant oligopolies: Google, Apple, Facebook, and Amazon. In this sense, the oligopoly of technological support is the new power that controls content dispersion. Although it is not an oligopoly of messages because the content is co-produced by users, it serves to control the information from the filters of algorithms and technological dominance. Given this vision, authors critical of network mediations, such as Umberto Eco (1965), regain meaning because the controls are still present within the new media strands. The problem is not to ask whether, conveyed by the same narrative schema, different ideological "contents" can have different effects. Instead, it is this: an iterative system becomes and remains as such only insofar as it sustains and expresses semantic references that are, in turn, deprived of development. In other words, a narrative structure expresses a world, but we realize this even more by revealing that the world has the same structure as the configuration that expressed it (Eco 1965, 271).

In Eco's theory, the mass media are perceived by two antagonistic concepts: apocalyptic vs. integrated. On one side, the apocalyptic ones, those who saw mass culture as a reference to decadence, but on the other, the integrated ones, those who see the phenomenon as the ascension of all social strata with the freedom to express themselves. Conflicts between pessimists and optimists. The apocalyptic made his audience consider themselves above average intellectually. According to Eco, he was Superman because he was above the rest of the mass media. The integrated, however, invited the reader to passivity by accepting the uncritical consumption of the products of mass culture (Eco 1965).

The big difference concerning other technological revolutions of the past is that, in the current one, the data are the information shaped by the new medium, the computer. As Castells puts it, "it is technology to act on information, not just information to act on technology (Castells 1992, 78)." Compared to Eco's analysis, it was necessary to understand the workings of culture and technology and know the context in which one works. It meant recognizing that the production of content

is controlled by economic interests. Machine learning is embedded in a paradigm that seeks, with all efforts, to reduce it into algorithms that can be managed by a few groups. A capacity bound by profit interests, as the apocalyptic of others would condemn. In the next section, we will examine journalists' limitations in understanding and implementing automation.

2.2.1 Journalists' limitations with automation understanding

When we talk about multidisciplinary potentials, the difficulties of understanding what is behind algorithms are an essential point, as well as the journalist's own inability to assimilate statistics for their production. It is common sense that journalists generally do not have a friendly relationship with mathematics, even though this use is made since data are necessary to give credibility to the news. Conventional reporting practices, such as collecting and processing the data made available in digital databases (Gray, Chambers, and Bounegru 2012), require specializations that professionals and scholars in the social sciences and humanities usually do not have, such as statistics and mining.

In a survey through interviews with prominent professionals in their fields in Rio Grande do Sul, Brazil, Marco Gehelen (2016) found that 73% of journalism students and 60% of conventional journalists in various newsrooms claim not to like mathematics. However, 82% of professionals said they usually make news and/or reports using numbers (Gehelen, 2016). Pereira et al. (2015) conducted a similar study to understand mathematical errors in Portuguese daily newspapers in Portugal. According to the authors, the results revealed the existence of mathematical errors in almost half of the texts analyzed from *Correio da Manhã* (45%), in about 35% of the texts from *Público*, and in about 18% of the texts from *Jornal de Notícias*. Still, as they themselves compliment: "The most frequent errors are of the subjective type (S. Pereira, Azevedo, and Machiavelo 2015, 516)."

According to Meyer (1991), journalists routinely misinterpret statistical data and tend to err towards overinterpreting coverage, for example. In another vein, Cohn and Cope (2011) complement that the press was influenced by charges to

launch scoops, which contributes to definitive publications of provisional numbers. All cast a view of a need to reinvent themselves and foresee possible disruptions of the traditional work structure, as argued by USA Today's chief editor for special projects, Julia D. Wallace:

I imagine the key point here is going to be the need to learn a lot more things to get to be a journalist in the future. It's no longer enough to have writing skills, good sources, and tons of energy. Writers and directors will need to know something about statistics and the computer processing of information (Wallace, 1991, cited by Meyer, 1991, p.45).

In a growing wave and due to these scenarios, Data Journalism has become increasingly celebrated and advocated, precisely because of the expansion of computers and other tools in newsrooms, especially about the incorporation of interfacing with computer science researchers in the reporting, editing, publishing, and broadcasting of content (Träsel 2014). The use, precisely, serves as a parameter to expand investigative journalism and act on the credibility of the content.

2.2.2 Artificial Intelligence Definitions

As defined by Nilsson □, Artificial Intelligence applications encompass some primary goals of reproducing human ability, such as search, planning, knowledge representation, perception, ability to move objects (robotics), natural language programming (NLP), and learning. The applications also usually correspond to subfields within these objectives, which are vehicle automation (drones, autonomous cars), medical diagnosis, artistic creation (poetry), applied mathematics, games (like chess), internet search (like Google), online assistants (like SIRI), photo recognition, SPAM identification in emails, judicial prediction, online advertising through segmentation, etc. (Nilsson, 2009).

In journalism, when evaluating the cautions and the side to be fought against in the use of technology, a more explicit position emerges on how to act with the favorable horizons that are on the upside and that are born through journalism automation tools. The newsrooms are already adopting precise techniques and principles for the use of emerging technologies around the world. To give some

examples, the Washington Post used an automated writing tool to cover the 2016 election race. During the same presidential campaign, The New York Times introduced a newsbot that generates news alerts on Facebook messenger (Howard, Woolley, and Calo 2018).

Reducing the human effort in tasks that journalists find difficult or even incomprehensible, thereby bringing the best quality news to the public, is to be celebrated. Marconi and Siegman simplify the prominence of artificial intelligence in journalism into five subareas, which are machine learning, natural language programming (NLP), speech transcription from written text to spoken speech and from spoken speech to written text, computer vision, and robotics (F. Marconi and Siegman 2017). The connection between the fields of communication and computing, as partners in the same challenge and similar ethical dilemmas, makes room for the computer science journalist in a vital role of understanding the dilemmas and the most appropriate ways to communicate and contribute to computer scientists somewhat. As they point out, “computational journalists are principally journalists, but they understand how artificial intelligence works and how AI can be used to augment their own journalism. Perhaps most importantly, they know how to communicate well with data scientists” (Marconi e Siegman, 2013, p. 5).

To support journalists in this capacity, this section will delve into the main contributions of artificial intelligence to journalism today. These contributions come at a time when the press and media companies are seeking new solutions to the challenges posed by the digital revolution. As we have described, Artificial Intelligence is the branch of computer science that develops algorithms mimicking human reasoning and capabilities to make decisions, solve puzzles, and perceive novel solutions. Machine learning, a subfield of AI, has seen significant recent advances. It focuses on identifying rules and patterns within large data sets, known as big data. These rules are established in generic algorithms that define information about a composite of data, eliminating the need for specific programming code to solve each problem (N. Nilsson 2014). Instead, the algorithm derives its logic from the data (Figure 10). In simple terms, rather than writing code for every action, the generic algorithm builds its logic based on the data input (Geitgey 2017).

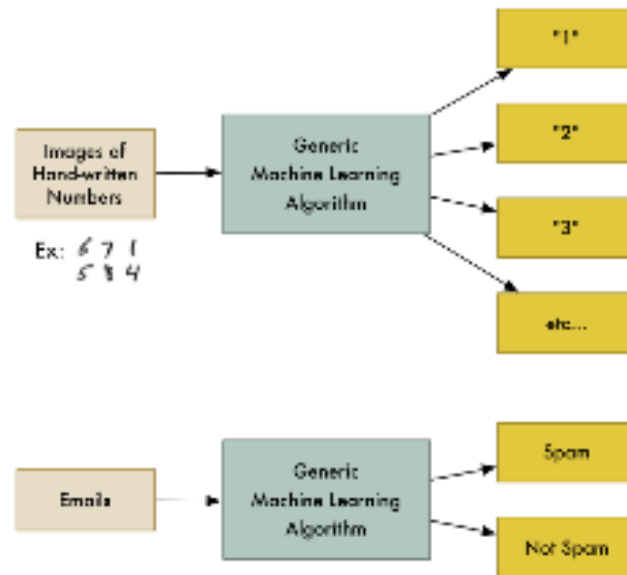


Figure 10. Exemplification of algorithm generalization (Geitgey, 2017)

For Nilsson (2014), learning algorithms present a logic usually designed by three basic steps: identifying a pattern or model, predicting it, and evaluating how well the model works. Each step uses two possible paths. In the first step (Pattern), you can follow the path of inference or prediction (infer or predict). In the second (Predict), the path of hit or miss (arrow or loss). Moreover, finally (Learn), of training or learning (train or learn). For each model assumed, these three steps are calculated and trained in a matrix that receives the weights for each specificity (features), which assume one or another validation, according to models whose efficiency in the result is directly related to how many million data are imputed.

To summarise, machine learning can be classified into three types: supervised, unsupervised, and reinforcement (N. Nilsson 2014). Supervised learning encompasses many applications with more expressive results because it addresses well-understood problems. Supervised learning problems are often divided into regression and classification tasks (Figure 11). In a regression problem, predictions are made based on continuous numerical data, such as predicting a house's price based on its size. In contrast, a classification algorithm categorizes information into discrete groups. For instance, the same classification algorithm can identify a cancerous tumor by its size or classify an email as SPAM. The algorithm adapts to different classification tasks based on the input data (Geitgey 2017). In journalism,

supervised learning can analyse large datasets to identify trends and generate reports. For example, algorithms can classify news articles by topic, sentiment, or source reliability, aiding journalists in managing and curating content more effectively (Diakopoulos 2019).

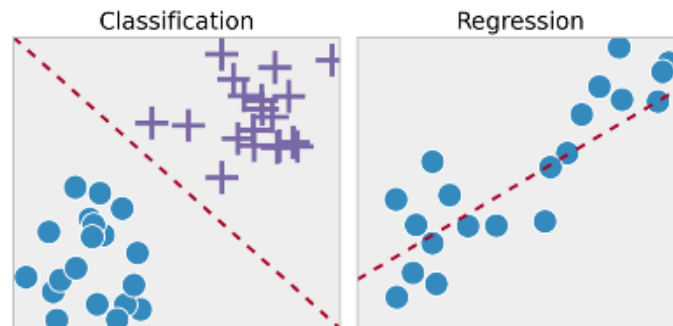


Figure 11. Parameters for differentiating classification and regression algorithms (Geitgey, 2017)

Unsupervised learning deals with unlabeled data, aiming to identify patterns or groupings within the data. It does not rely on predefined outcomes. This type of learning is used for tasks such as clustering and association. For example, in customer segmentation, in the same previous scenario about the prices of the houses, unsupervised learning can group customers based on purchasing behavior without prior labels, revealing natural clusters and associations (Figure 12). According to Geitgey (2017), this type of learning seeks relationships within the data, often identifying clusters and trends without external feedback. Yann LeCun, Facebook's CEO, notes, "while supervised learning is the icing on the cake, unsupervised is the cake as a whole (LeCun 2016)." In journalism, unsupervised learning can be applied to uncover hidden patterns in data, such as identifying emerging trends in social media discussions or clustering similar news stories. This can help news organizations quickly identify and respond to new topics of interest or public concerns (Broussard 2015).

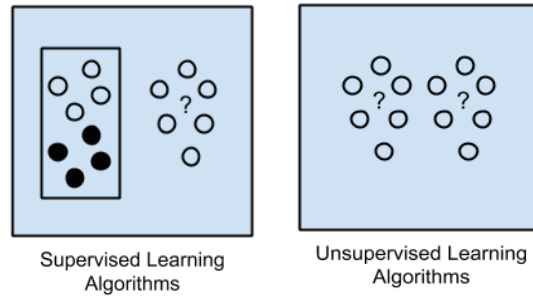


Figure 12. Differentiating supervised and unsupervised learning (LeCun, 2016)

Reinforcement learning involves training algorithms through a system of rewards and penalties. It is used in dynamic environments where an agent learns to achieve a goal by interacting with its surroundings and receiving feedback based on its actions. This approach, rooted in behaviorism theory by Skinner, uses rewards and punishments to reinforce behavior (Baum 2018). For example, reinforcement learning is used in robotics for navigation and game playing, where an agent learns strategies by receiving rewards for winning and penalties for losing. It considers uncertainty and adapts based on trial and error. In journalism, reinforcement learning can be used to personalize content delivery. For example, news recommendation systems can use reinforcement learning to improve the relevance of articles continuously suggested to readers based on their interactions with previous content (Letham et al. 2015).

Another fundamental concept is the distinction between general AI, or strong AI, and task-specific AI, or weak AI:

Strong AI is defined as intelligence that works across multiple application domains. It is an intelligence similar to that of humans, in that it is not specific to a task but rather can function in situations and contexts that are completely new. Weak AI is intelligence that works only in a narrow set of applications. It is the AI of today, and it includes the intelligence that drives autonomous vehicles, manufacturing robots, computer vision (Denton, Chintala, and Fergus 2015), and recommender systems (ResnickP 1997). Weak AI also includes the algorithms that have become famous for beating humans at various games, such as chess (M. Campbell, Hoane Jr, and Hsu 2002), Jeopardy! (High 2012), and Go (Silver et al. 2016), although the ability of some of these systems to learn by playing against themselves makes them quite versatile (Hidalgo et al. 2020, 15).

When advancing in the understanding of Machine Learning, deep learning has been highlighted in the automation of communication and journalism practices. This generalization model is a branch that uses generalization algorithms to imitate

somehow the procedure done by the human brain (Werbos 2006). Each layer in the network is a simple algorithm containing an activation function. Advances in neuroscience are present in interpreting these data, and whose patterns are represented in a nervous system, which, like a neural coding, establishes stimuli associated with the brain (Olshausen and Field 1996).

Artificial intelligence (AI) has its historical origins intertwined with the exigencies of World War II. In their pursuit of breaking German encryption, the United Kingdom and the United States enlisted the services of numerous mathematicians (Crawford 2021). Among these notable mathematicians was Claude Shannon, whose contributions proved seminal. Shannon discerned certain patterns within encrypted texts, specifically the sequential order and frequency of letters in the alphabet. He astutely observed that letters follow each other in a particular sequence and exhibit distinct frequencies. For instance, a noun typically succeeds an article; the letter 'e' is the most commonly occurring letter in the alphabet (Soni and Goodman 2017). To formalize these observations, Shannon developed a statistical method, which he referred to as a "language model." Essentially, a language model provides a probability distribution over a predefined vocabulary based on a given collection of words. This distribution reflects the likelihood of encountering these words in natural language sequences (Guizzo 2003).

Shannon's work on language modeling marked a pivotal step toward the development of AI. Concurrently, in 1951, two other mathematicians devised the concept of the artificial neuron. They demonstrated that by connecting multiple artificial neurons, a neural network could be constructed, capable of autonomously learning and modeling sequences of words. These neural networks paved the way for the creation of extensive language models (Benko and Lányi 2009). Remarkably, it was discovered that when these language models were scaled up significantly, they exhibited a form of intelligence.

Until a few years ago, and still in the early days of artificial intelligence applications, deep learning models had been overlooked. However, with the advent of the vast amounts of data available in digital life, their relevance has resurged and reached new heights. Significant advances come from reading

unstructured data, which include text, images, and videos (Goodfellow, Bengio, and Courville 2016). These unstructured data, however, remain challenging to interpret, with issues of unclear interpretation and causal inference, often referred to as the Blackbox problem.

Identifying and distinguishing images with numerical ratings representing pixels is not straightforward for a computer. Each small segment of an image can represent various textures, such as those of an animal, a car, or a person. Researchers Goodfellow, Bengio, and Courville (2016) at MIT explain that the great differentiator of deep learning lies in its ability to break information into numerous mappings, which end up categorizing new layers into subsequent sequences that help categorize those images more easily (Figure 13). For this process, the greater the quantity of information about the same image, the more accurate the reading. To clarify, they refer to an image of a woman. In decoding this data, the pixels are categorized into slices that progressively reveal the abstractions of the image. Finally, the subdivisions acquired from multiple images can distinguish and classify the images effectively (Goodfellow et al., 2016).

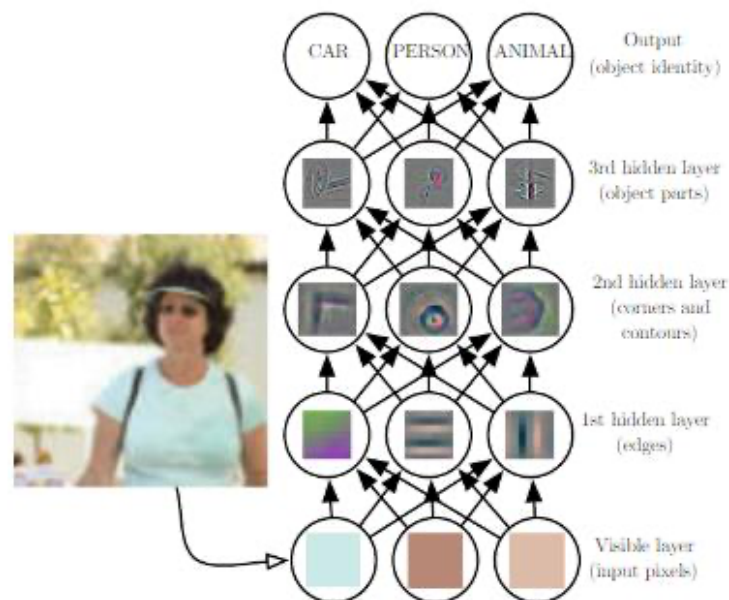


Figure 13. Illustration of a Deep Learning model in the layers (Goodfellow et al., 2016)

In the context of journalism and communication, Natural Language Processing (NLP) and generative media have become crucial elements powered by deep learning. NLP enables machines to understand, interpret, and generate human language in a valuable way. This includes machine translation, sentiment analysis,

summarization, and entity recognition. For example, NLP algorithms can analyze vast amounts of text data from news articles to identify key topics, trends, and sentiments, helping journalists to gather insights quickly and produce timely reports (Diakopoulos 2019).

Generative media, on the other hand, refers to the creation of content by AI systems. This includes text, images, and even videos generated by algorithms. Generative models, such as GPT-3, have been used to produce coherent articles, generate social media posts, and even create conversational agents that can engage with users in a human-like manner (T. Brown et al. 2020). In journalism, generative media can automate routine writing tasks, generate personalized content for readers, and assist in investigative reporting by generating plausible scenarios and narratives based on data (Carlson 2015). Deepfake technology, though controversial examples presented in the previous section, showcases the potential of generative media by manipulating video footage to create realistic portrayals of events that never occurred. We'll explain how these areas work in more detail.

2.2.3 Natural Language Processing (NLP) Applications

Text learning, Natural Language Processing (NLP), has seen remarkable advancements, particularly in translation applications, document reading, and measuring emotions in expressions. As Bird et al. clarify, "natural language" refers to languages used in everyday communication, such as Portuguese and English, as opposed to programming and mathematical languages, which have evolved over the Years (Clerwall 2014; Davis 2014). In its analysis, NLP also incorporates a constant transformation and readaptation. For this reason, this subdivision of artificial intelligence is concerned with using applications and models to define resources of statistical probability in sequences of words, characters, or bytes of a natural language.

NLP uses a base of deep machine learning algorithms and another floor of algorithms developed primarily for its application. Natural language forms included in the NLP: image guides, search engines, spoken language, word processing, etc. Phones and tablets already support predictive text and handwriting recognition; web search engines provide access to information through unstructured data;

automatic translation allows us to translate texts written in Chinese into Portuguese; Text analysis enables us to detect feelings in Facebook comments (Bird, Klein, and Loper 2009). Language processing plays a fundamental role in technological society by connecting human communication with algorithms more naturally and accurately.

One of the first steps in any NLP is to distinguish the language in which the document is represented (Bird, Klein, and Loper 2009). The linguistic Corpus from Corpora English represents the set of texts and oralities in each language and serves as a basis for analysis. Despite their basic applicability, computational methods are fundamental for this first screening. These grammatical concepts are processed into more abstract structures to model language and enable machine understanding, which reduces it to the most relevant content. The first language models were already inspired by fixed-length sequences called n-grams.

As Jurafsky and Martin clarify (2012), an n-gram is a sequence of n signs: Models that assign probabilities to sequences of words are called language models or LMS. The simplest model that assigns probabilities to sentences and sequences of words (the n-gram). An n-gram is a sequence of N words: a 2-gram (or bigram) is a two-word sequence of words like "please turn," "turn you're," or "your homework," and a 3-gram (or trigram) is a three-word sequence of words like " (Jurafsky and Manning 2012, 38). Tokenization involves the transformation of uppercase letters to lowercase, removing characters, removing HTML/Javascript/CSS tags, etc. An example of tokenized in the Figure 14 as follows:

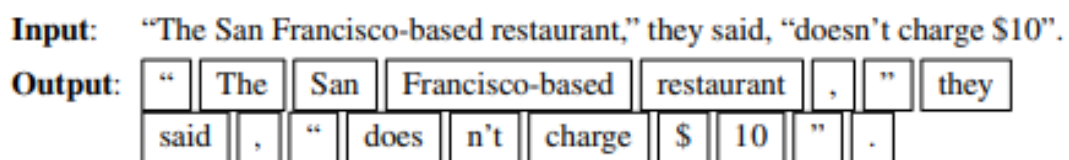


Figure 14. Example of LDC tokenization, which in this case uses the abbreviation

As for the morphology of words, Ontology is a form of classification that helps to promote semantic analysis and correlation (Figure 15). As Heravi and McGinnis

define it, through ontology, notes are taken more broadly and generally about the organization of the text, as described:

Ontology-based Information Extraction techniques will assist in exploiting the hierarchical structure of an ontology to inform the entity extraction task. Three main stages are involved within this task: (a) detection of entities in the text (including specification of entity types); (b) search for potential matches in Linked Open Data; and (c) entity ranking to find the most appropriate entity to link the mention to (Heravi and McGinnis 2015, 136).

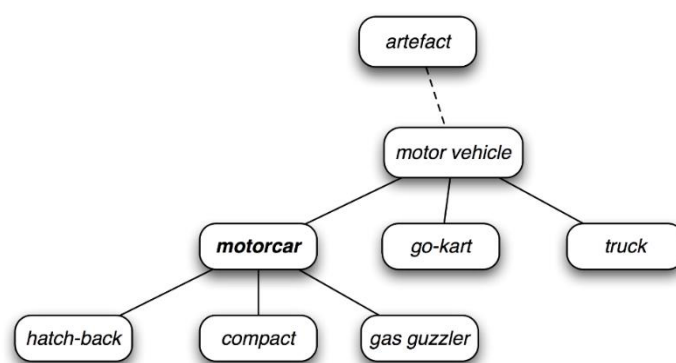


Figure 15. Fragment of a hierarchical ontology (Bird, Klein, and Loper 2009)

A practical example of NLP applied in journalism is the use of automated news generation systems. These systems utilise NLP to automatically generate news articles based on data inputs, such as financial reports, sports results, or election outcomes. For instance, The Associated Press (AP) employs an automated content generation platform called Wordsmith (S. JohnSon and Ensslin 2006), developed by Automated Insights, to produce financial earnings reports. This system analyses structured data from financial reports and converts it into coherent news articles, allowing journalists to focus on more complex stories and investigative work.

Bringing examples of NLP in the Portuguese context, Nunes et al. (2024) developed the Text2Story Lusa dataset, consisting of European Portuguese news articles, to aid narrative analysis in linguistics and computer science. The primary dataset includes 357 news articles, while a subset of 117 articles features dense manual annotations capturing morphosyntactic and semantic details. These resources, designed to facilitate narrative extraction and other NLP tasks, address the scarcity of freely available annotated datasets in this domain. The project aims to support various applications, including developing algorithms for automated narrative extraction, analysis of news article structures, and improvements in natural language understanding. Researchers can use the dataset to train machine

learning models to understand better narrative elements, such as plot progression, character development, and event causality. Additionally, the dataset can aid in the creation of tools for journalists and content creators, helping them to analyze and generate high-quality narrative content (Nunes et al. 2024).

2.2.4 Generative Media: Advances in Language and Vision Models

In discussing language models, it is imperative to mention GPT-3 (Generative Pre-trained Transformer-3), a significant milestone in the field (McGuffie and Newhouse 2020). Launched in 2020, this model can generate texts with coherent language structures, such as essays, long texts, translations, and even computer code. Developed by OpenAI, a San Francisco-based research lab, GPT-3 is a "big data language model" using deep learning trained on texts from thousands of books and most of the internet to string together words and phrases. Notably, the English Wikipedia comprises only 3% of the training dataset for this model, highlighting its extensive data usage (Dale 2021). The texts generated by the system are often indistinguishable from those written by humans, particularly in shorter formats.

From 2018 to 2020, the conventional approach involved fine-tuning models through limited supervised learning, customizing them for specific tasks. Recently, however, researchers discovered that the largest models, exemplified by GPT-3, exhibit exceptional performance on novel tasks with minimal input. Even a simple human language description or a few examples can prompt these models to excel in tasks for which they were not specifically trained (T. Brown et al. 2020).

In conjunction with GPT-3, BERT (Bidirectional Encoder Representations from Transformers) is another significant pre-trained neural network-based algorithm for NLP functionalities. Developed by Google in 2018, BERT improves the contextual understanding of unlabeled text by learning to predict the text that might come before and after. According to Koroteev, one of BERT's functionalities is distinguishing words with multiple meanings based on their surrounding context, thus aiding in discerning the meaning of words in search queries. The model is an unsupervised language representation that uses a plain text corpus and converts words into numbers (Koroteev 2021).

Over the past five years, large language models have made significant strides, achieving a level of intelligence exemplified by models like ChatGPT. Significant developments in text automation include the release of AI-based conversational models such as Google's Bard and OpenAI's ChatGPT. Bard, powered by the Language Model for Dialogue Applications (LaMDA), and ChatGPT, derived from GPT-3.5, are vying for market leadership. LaMDA is a transformer-based neural language system pre-trained on online chat data (Rahaman et al. 2023a). ChatGPT, released in November 2022, uses a reinforcement learning model with human feedback, offering high accuracy and efficiency in a short period (Rahaman et al., 2023). While Bard was developed for chatbot use, ChatGPT can produce text in more varied contexts and generate responses that enhance discussion authenticity (Kitamura 2023). Bard's ability to produce text in various tones and voices provides flexibility as a content-generation tool (Rahaman et al., 2023).

Moreover, computer vision models such as DALL-E and SORA exemplify the cutting-edge capabilities of generative AI in producing diverse and realistic images. DALL-E, developed by OpenAI, generates images from textual descriptions using a neural network trained on a vast dataset of images and their corresponding captions. This model exemplifies how generative AI can create novel visuals by interpreting and synthesizing elements from the input text, thereby pushing the boundaries of image generation and creativity (Y. Liu et al. 2024).

According to Allaham and Diakopoulos (2024), integrating Large Language Models (LLMs) into newsroom operations presents a crucial decision point: whether to invest in developing proprietary LLMs or employ existing models through techniques like prompt engineering, fine-tuning, and retrieval-augmented generation (RAG). For example, BloombergGPT, a specialized LLM developed for financial data analysis, required an estimated \$1 million solely for training, excluding deployment infrastructure costs (S. Wu et al. 2023). Developing LLMs from scratch involves significant financial and resource commitments. In contrast, leveraging third-party models offers a cost-effective and flexible alternative. These models, accessible via API endpoints, reduce infrastructure expenses and allow rapid adaptation to new technological advancements (Allaham and Diakopoulos 2024).

Allaham and Diakopoulos (2024) assert that retrieval-augmented generation (RAG) is particularly advantageous for boosting the productivity of news organizations using large language models (LLMs). Employing methodologies such as prompt engineering (Y. Wu and Hu 2023), RAG, and fine-tuning is essential. Prompt engineering, which involves creating precise queries to extract specific information from LLMs, requires significant domain expertise to enhance the efficiency of the information extraction process. RAG goes a step further by integrating the LLMs' inherent knowledge with real-time or specialized external data, thereby increasing the precision and relevance of the results. Fine-tuning allows for customization of pre-trained LLMs to suit specific organizational needs and styles, diminishing the need for extensive prompt engineering. Together, these strategies promote the effective adoption of third-party LLMs in news organizations, ensuring both cost-effectiveness and adaptability to the fast-changing landscape of news production (Allaham and Diakopoulos 2024).

Nonetheless, language models are not infallible. They occasionally grapple with uncertainty, leading to "hallucinations" or the inadvertent production of false information (Azamfirei, Kudchadkar, and Fackler 2023). These hallucinations are not intentional deceit; they result from the model's inherent limitations. A case in point is a lawyer seeking relevant legal precedents and receiving a list from ChatGPT. Subsequently, it emerges that these cited cases are entirely fabricated, illustrating the model's capacity for unintentional falsehoods. While hallucinations are not malevolent, they do present a significant issue (Alkaissi and McFarlane 2023). Despite their potential, language models cannot possess exhaustive knowledge of the world and represent some highly risky uses. Nevertheless, they earnestly endeavor to maintain continuity in the given sequences. In this context, they reflect human behavior, where individuals may embellish their knowledge in areas of uncertainty (J. Wang et al. 2023).

The predicament of hallucinations becomes acutely concerning when producing content that necessitates factual accuracy, such as news articles. Reliance on language models for copywriting demands a discerning approach. Efforts are ongoing to mitigate the likelihood of hallucinations, exemplified by technologies like retrieval augmented generation (Lewis et al. 2020). This

approach could enlist the language model to conduct preliminary research, akin to a journalist, and references a dataset, perhaps archives from a fact-checking with human supervising, to collate trustworthy information and generate content with a diminished susceptibility to hallucination. The potential avenue involves linking language models with well-organized repositories of fact-checking knowledge, either in the form of knowledge graph neural networks or vast reservoirs of text that can be accessed in real time. Continuing to the next section, we will delve into the mechanisms and applications of automatic fact-checking.

2.3 Automatic Fact-checking

Automated fact-checking represents a facet of investigative journalism with limited applicability in unique scenarios. Its most immediate benefits lie in automating routine tasks such as data extraction from varied documents and probabilistic record linkage across databases (Stray 2021). Essentially, automation's primary value is handling repetitive tasks like claim mining. However, the perceived potential of automated fact-checking can lead to misunderstandings about its capabilities, resulting in unrealistic expectations—challenges like those faced in other industries (Davenport and Ronanki 2018). Recognizing that disinformation is predominantly a social and political issue that requires more than just technical solutions is crucial.

The growth in the automation of written and spoken language indicates a new perspective for producing content, offering possibilities to address challenges using machine learning, synthetic media, and efficient communication fact-checking methodologies. Initiatives for manual fact-checking, such as FactCheck.org, Snopes, PolitiFact, and Full Fact, are also leaders in automated fact-checking research. Since misinformation has become a significant global concern, tech companies and international agencies have increasingly engaged in this area. These organizations are both beneficiaries and leaders in automated fact-checking research, which includes system development (Hassan, Li, and Tremayne 2015), and evaluation campaigns (Barrón-Cedeño et al. 2020). However, automated methods have credibility issues (Arnold 2020). The role of technological artifacts in

sociotechnical systems is essential, and when designing new technology, it is crucial to consider stakeholders such as data subjects, data actors, and model owners (Van de Poel 2020). Understanding who will use the technology is vital for explainability, as users may interpret explanations differently (Schuff et al. 2022).

Studies from Full Fact Org (DeFleur 2013) present the automatic phases in four subtasks (Figure 16). The leading subtask is monitoring, which involves extracting possible claims. Later, in the spotting subtask, the most numerous claims are selected. The checking subtask analyses the share corresponding to its fairness, and the final subtask consists of publishing, which refers to presenting the results in a human-readable format and highlights. These four phases will be used as eligibility criteria in our systematic review.



Figure 16. Four phases required to perform Fact-checking (Saquete et al. 2020)

Considering these possibilities, we will examine how fact-checking has been used with automation concerning health issues during the COVID-19 pandemic. This part of our work intends to investigate the extension of fact-checking for communicating health risks, addressing the urgency to identify and promote accurate alerts to the population regarding a global infodemic. Limiting the impact of these adverse developments has become a significant focus for journalists, social media companies, and regulatory authorities.

Automatic fact-checking systems can enhance manual fact-checking in these steps: (a) finding claims from the significant information ecosystem; (b) finding previously fact-checked claims; (c) finding supporting evidence, translating, and summarizing relevant posts, articles, and documents if needed; and (d) detecting claims that are spreading rapidly to slow them down (Nakov et al. 2021). Fact-checking is not an effortless process since it requires many steps (Figure 17) that range from scouring media to spotting check-worthy claims to concluding whether the claim is valid, partially accurate, false, misleading, or not possible to judge.

Below, we will compare two initiatives applying some COVID-19 informational data to be checked.

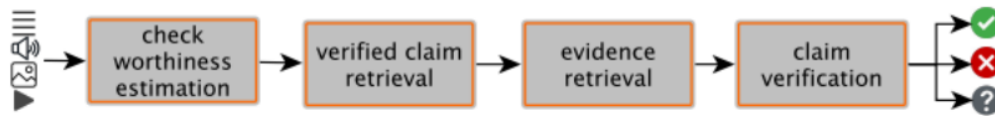


Figure 17. Prototypical fact-checking pipeline (Barrón-Cedeño et al. 2020)

To monitor such a range of available information, fact-checkers often use various technologies, such as automatic speech recognition, news alerts, and translation tools, which typically depend on underlying AI technologies (Arnold 2020). Since fake news spreads six times faster than real news (Owen 2017), technologies could allow fact-checking organizations to be quicker and provide more far-reaching coverage than manual verification. They also point out that automated claims verification seems like AI's ultimate application to fact-checking because many claims are not simply 'false' or 'true' but 'partially false' without extra context (Miller 2020). One key role of professional fact-checkers is to help their audience understand the issue, despite showing a binary label. In this study, we will systematically analyze the scientific research published during the pandemic to understand how these concepts are documented.

As discussed earlier, the COVID-19 pandemic gave rise to an infodemic, characterized by a massive influx of health-related content being rapidly disseminated through social media channels. Some researchers have identified discrepancies between fake and real news in this scenario. One example is the study by Mu et al. (2023), which conducted a large-scale computational analysis of more than 242 million tweets. The primary objective was to examine and contrast the specific attributes of COVID-19 misinformation and accurate COVID-19 information. The study focused on four key aspects: 1) the distribution of topics, 2) the real-time status of tweets, 3) language analysis, and 4) the dissemination patterns over time. Additionally, the research team curated a dataset for classifying COVID-19 misinformation, further aiding in the understanding and management of such content. Some graphs (Figure 18) clearly provide valuable insights into the

spreading power of both types of tweets over time. In the context of AI applications in journalism, we now transition to exploring AI tools specifically designed for detecting disinformation in the digital age.



Figure 18. The first graph represents the spread of misinformation, and the other showcases the spread of non-misinformation content. The data showcases the average number of times the source tweets are retweeted or quoted within 4-hour intervals (Mu et al. 2023)

2.3.1 AI Tools for Detecting Disinformation in the Digital Age

While fact-checking remains a primarily human-led endeavor, the sheer volume of misinformation circulating on the internet exceeds the capacity of manual processes alone. This section explores a selection of AI-powered tools that can assist reporters and fact-checkers in identifying and addressing disinformation across different mediums, including text, images, videos, and bot activity.

One notable tool in the fight against disinformation is Exorde Network. Developed as a global information processing platform, Exorde analyzes the spread and virality of information on the internet. By examining URL submissions and processing the sentiment of claims, Exorde provides insights into people's reactions to specific content. Furthermore, it archives this information for public accessibility,

facilitating a more comprehensive understanding of disinformation dynamics ("Exorde" 2023). FactCheck.org, a project hosted by the Annenberg Public Policy Center at the University of Pennsylvania, focuses on verifying claims made by American politicians at various levels of government. With the assistance of Ask FactCheck and Ask SciCheck, journalists can have their information fact-checked by the FactCheck.org team, contributing to the accuracy and reliability of their reports (Jackson 2012).

Another tool, Factiveuse, offers an AI-driven approach to expedite the fact-checking process. Based on research conducted at the University of Stavanger, Factiveuse developed an AI Editor capable of identifying disinformation in content and directing users to credible sources. Factiveuse also provides a fact-checking database called FactiSearch and a context-providing tool known as Microfacts, which explains jargon and complex concepts ("Factiveuse" 2023). To enhance the visibility of fact-checked stories on search engines and social media platforms, ClaimReview, a tagging system developed in partnership with Google, promotes verified information by displaying fact-checked statements alongside a note indicating their validity ("ClaimReview" 2023). This initiative aims to combat the rapid spread of disinformation by prominently presenting fact-checked content to users.

In video verification, the InVID-WeVerify verification plugin offers a comprehensive set of tools. With features such as Image Content Analysis, Optical Character Recognition in Images, and the ability to compare content against a Database of Known Fakes, journalists can fact-check visual content effectively (Nygren et al. 2021). Truepic offers a solution for validating the authenticity of visual content. Truepic Display provides information about the source and history of any media file, while Truepic Lens offers metadata analysis. By leveraging these tools, journalists can verify the origin and credibility of visual content more effectively (Omezi and Jahankhani 2020).

Detecting bot activity is crucial in understanding the dissemination and amplification of disinformation. The Botometer tool, created by the Observatory on social media (OSoMe) at Indiana University, assigns scores to Twitter accounts, distinguishing between human accounts and bot accounts. Botometer further

categorizes bot accounts into types such as Echo-chamber, Fake follower, Financial, Self-declared, Spammer, and Other, aiding in the identification of automated disinformation campaigns (Yang et al. 2019). Hoaxy is another valuable tool that tracks the spread of information originating from low-quality, human-generated websites. Journalists can utilize Hoaxy to trace the virtual path of articles and assess the likelihood of bot-assisted dissemination (Shao et al. 2016).

Beyond the aforementioned categories, several additional tools are worth considering. Media Bias/Fact Check (MBFC) offers a platform for verifying fact-checkers and news websites ("Media Bias Fact Check" 2019), rating them based on bias, political affiliation, and the authenticity of fact-checked information. With reports covering more than 6,300 media sources, MBFC provides journalists with valuable insights into the credibility and trustworthiness of different media outlets. The Hamilton 2.0 Dashboard ("Hamilton 2.0" 2023), created by the Alliance for Securing Democracy at the German Marshall Fund of the USA, focuses on analyzing media messages and narratives promoted by the Russian, Chinese, and Iranian governments. These tools provide invaluable support in identifying, verifying, and debunking false information, empowering journalists to uphold the integrity of their reporting.

2.3.2 Publication selection criteria and methods

Against this background of existing solutions, our project aims to carry out a systematic review to find answers to the hypothesis that there are possibilities to complement the theme significantly. To obtain data and find a research focus, we will adopt a systematic review to help us identify the most effective ways to contribute to this vast topic. We will use a systematic literature review (SLR), following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al. 2009a). The guidelines suggest a checklist of 27 items under four phases in the systematic review process. We've used three databases (ACM Digital Library, Scopus, Google Scholar, and Web of science) to extract the most relevant literature as data sources from 15-nov. 2022 to 16-nov. 2022. These databases have been selected because they are the ones that return the

most significant number of results in this area due to their coverage (Guerrero-Bote et al. 2021). The search strategy is encircled around the keywords search strings: ("Disinformation" OR "Health" OR "Pandemic") AND ("Automat*" OR "Fact-checking") AND ("journalism") AND ("Informed citizens"). It was applied to capture the studies published during the Covid-19 Pandemic from 2022 to 2020. The planning was done to discover potential contributions in journalism automation focused on the health infodemic.

The search included the eligibility criteria as studies discussing fact-checking related to health and studies presenting technological automation solutions under the four already submitted phases from fact-checking automation (monitoring, spotting, checking, and publishing). To narrow the search, we've applied the exclusion criteria of studies out of the Covid-19 years (2020 to 2022) and research strictly theoretical. The data collection process was planned to use PICOT criteria as a) population: pandemic, disinformation, health infodemic; b) intervention: automatic fact-checking, c) comparison: traditional journalism; d) outcome: informed citizens; and e) context: covid-19 pandemic, vaccination hoaxes. The flow diagram (Figure 19) presents the process followed for the Systematic review. Then, a reading of the full text was planned to extract evidence in the literature on automation and AI for fact-checking and a research area that could be better used according to the literature.

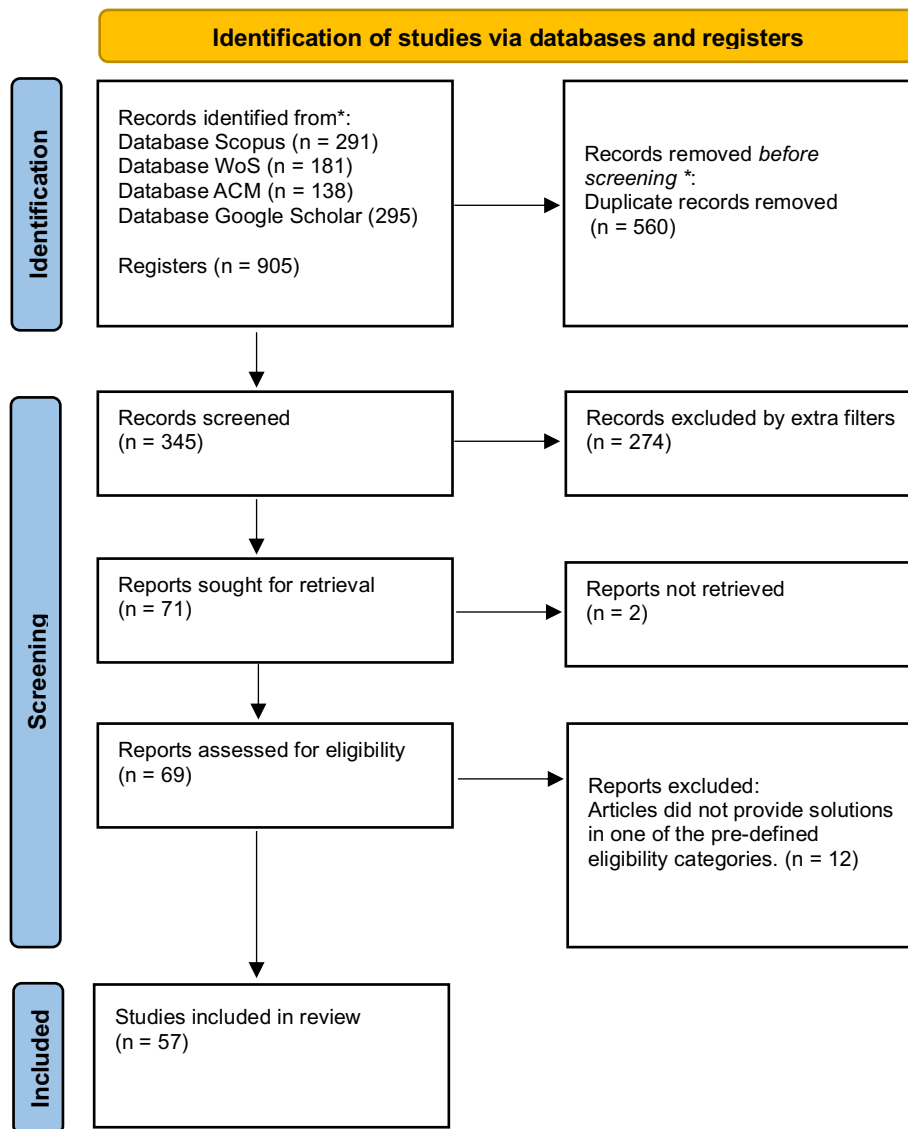


Figure 19. PRISMA 2020 flow diagram for automatic fact-checking new systematic reviews

First, we reviewed the title and the abstracts. Then, we applied the planned criteria presented before, including study screening, eligibility, and selected articles with variables about the title, year of publication, and details about the categories they could refer to. As illustrated, a total of 905 records were extracted by initial search query from five datasets: ACM Digital Library (138), Scopus (291), Google Scholar (295), and Web of Science (181). After assessing duplicate articles, using Parsifal software (Kitchenham and Charters 2007), we got 350 pieces of literature for further analysis. A review of titles, keywords, and abstracts has been conducted to include the literature containing automatic fact-checking tools content in the

health pandemic period. As a result, 270 studies were excluded, which were - publication date range (20), literature surveys (35), and context (215).

Seventy records were considered for full-text reading, and ten were excluded because they were strictly theoretical. Finally, 57 papers were selected to provide an overview of the automation technologies developed and used for fact-checking during the pandemic. To mitigate the risks of biases, a quality assessment checklist was done to appraise it. Table 3 question to be answered for each identified paper. The reason for extracting these answers from the existing literature is to provide a piece of overall knowledge about fields to be explored about the subject. This method also presents confidence in the body of evidence for an outcome of future research and prototype proposals.

Table 3. Quality Assessment Checklist

Question	Answer 1	Answer 2	Answer 3	Answer 4
Do the articles present automation solutions used in fact-checking that could prevent health disinformation?	Should report about monitoring automatic technologies (extraction from possible claims).	Should report about spotting automatic technologies (most many claims are selected).	Should report about checking automatic technologies (analyze the claim corresponding to its fairness).	Should report about publishing automatic technologies (AI to share the results).

2.3.3 Exploring the results and their correlations

A total of 904 publications were found in the period of 2020-2022, which present a wide range of scientific topics. Figure 20 shows the top 10 conferences in which the selected 57 publications were presented, thus comprising about 10% of all publications. The top two conferences with more publications were: Ceur Workshop Procedures (N=4), followed by Lecture Notes in Computer Science (N=3). Figure 20 shows that more four conferences also cited two sources each: Computational and Mathematical Organization Theory; MAD 2022 - Proceedings of the 1st International Workshop on Multimedia AI against Disinformation; NLP4IF 2021 - NLP for Internet Freedom: Censorship, Disinformation, and Propaganda - Proceedings of the 4th Workshop; Smart Innovation, Systems and Technologies. With this analysis, it is possible to know the sites that have dealt the most with this topic and selected technical content that discussed the topic of fact-checking automation in the pandemic period.

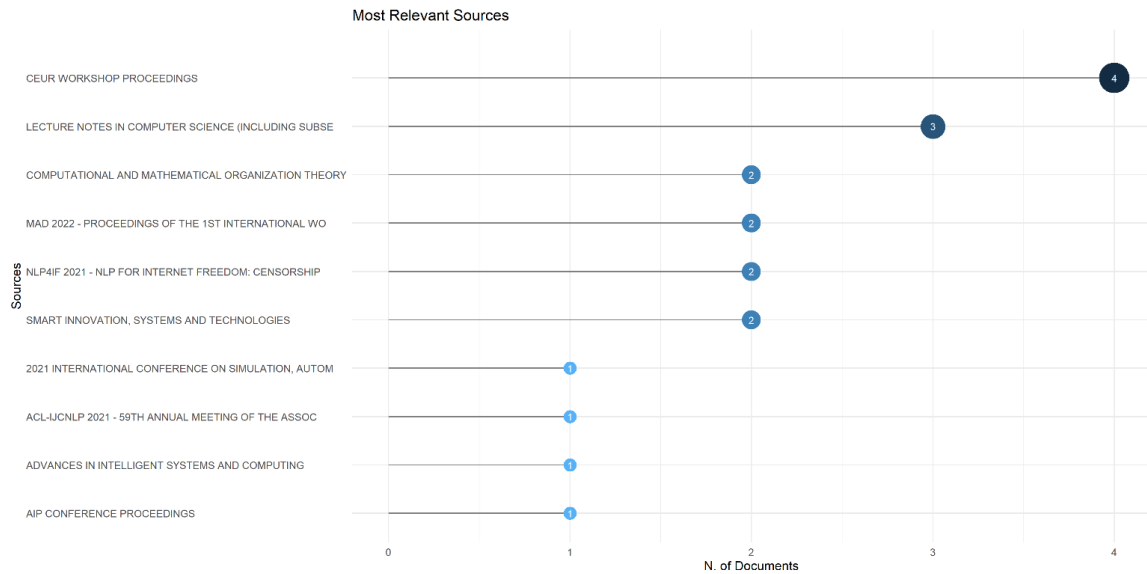


Figure 20. Top 10 conferences with the screened automatic fact-checking during the pandemic period

As for fact-checking and technologies, the meta-analysis answers for the questions are conducted to provide some target subjects for research or if the study evaluates the technology's impact during the pandemic period. They also present results, if any metrics are employed or an evaluation technique used to any specific phase from fact-checking, and which strategies and tools are employed in each. To bring quantitative data from the previously defined categories, we read and categorized them to indicate which concepts are more adequate to each one of them, even though there was no direct indication in the publications.

The classification intends to generate insights for future approaches, and currently we have no intention of going into technical detail about them. Most of the studies have checked as the object of the technological research (n=25), followed by spotting (n=20) and monitoring (n=18). Few studies addressed publishing technologies or even the theoretical approach for it (n=6). Table 4 shows the complete scattering of investigations according to the object of study. The studies are fundamentally analytical (n=18) and explanatory (n=6). Few cases make a comparative or descriptive investigation (n=3). The main information/data gathering technique used is experiments (n=19) and the collection of documents or

audio-visual/digital media materials (i.e., tweets, posts, YouTube videos, WhatsApp messages) (n=13).

Table 4. Fact-checking technologies divided by categories, references and quantity

CATEGORIES	AUTHORS*
Checking (N = 25)	(L. Liu et al. 2022); (Tashtoush et al. 2022); (Brand et al. 2022); (Aloshban 2020); (Nakov et al. 2021); (Botnevik, Sakariassen, and Setty 2020); (Elhadad, Li, and Gebali 2021); (Pathak, Srihari, and Natu 2021); (Hammouchi and Ghogho 2022); (Sarrouti et al. 2021); (Bonet-Jover et al. 2021); (Vedula & Parthasarathy, 2021); (Jiang et al. 2020); (J. Yao et al. 2021); (Charles, Ruback, and Oliveira 2022); (Babakar & Moy, 2021); (Karagiannis et al. 2019); (Shaar et al. 2021); (Springstein et al., 2021); (Smeros, Castillo, and Aberer 2021); (Simpson and Yang 2022); (Arnold, 2020); (Woloszyn et al. 2021); (Azevedo et al. 2021); (Trokhymovych and Saez-Trumper 2021)
Spotting (N = 20)	(Pogorelov et al. 2021); (Ng & Carley, 2021); (D.K. Sharma and Garg 2021); (Harrag and Djahli 2022);(Alhindi et al. 2021); (Babakar & Moy, 2021) ; (Wühlrl and Klinger 2021); (Mansour, Elsayed, and Al-Ali 2022); (Xin Wang, Chao, and Yu 2021); (Sarrouti et al. 2021); (Kotonya and Toni 2020); (Sawczyn et al. 2021); (Martín et al. 2022); (Saquete et al. 2020); (Carpuat, de Marneffe, and Meza-Ruiz 2022); (Zhu et al. 2021); (Shaar et al, 2020); (Nikolov et al. 2020); (Arnold, 2020);
Monitoring (N = 18)	(Dai et al. 2022); (Babakar and Moy 2016); (Casillas et al. 2022); (Savage 2021); (Nabožny et al. 2021); (Moffitt, King, and Carley 2021); (Neumann, De-Arteaga, and Fazelpour 2022); (Rakotoson et al. 2022); (Ding et al. 2022); (Srba et al. 2022); (Shaar et al. 2021); (Chen and Freire 2020); (Reis and Benevenuto 2021); (Arnold 2020); (Lanius, Weber, and MacKenzie Jr 2021); (Lahoti and Prasad 2022); (Villar Rodriguez et al. 2022); (Simpson and Yang 2022)
Publishing (N = 6)	(Vedula and Parthasarathy 2021); (Gruzd and Mai 2020); (Babakar & Moy, 2021); (Lanius, Weber, and MacKenzie Jr 2021); (Ng and Carley 2021); (Moravec, Kim, and Dennis 2020)

*Each author technique is described in detail in appendix A.

The theories used as a reference come mainly from the computer science area. Most of the research (n=43) adopts strategies for only one of the categories. In only one case, all the categories are presented, which is a theoretical description (Babakar & Moy, 2021), and in one of them, a tree of the categories is presented. In six cases tree categories are presented in the research. The “checking” category present techniques to detect falser news in China (Liu et. al., 2022), LSTM to automatically classify and identify Covid-19 fake news (Tashtoush et al., 2022;

Charles et al., 2022), veracity predictions (Brand et al., 2022), self-attention - ACT (Aloshban, 2020), browser applications (Botnevik et al., 2020), error analysis dataset (Pathak et al., 2021), Automatic bilingual annotation in Arabic/English (Elhadad et al., 2021), Covid-19 related framework (Hammouchi & Ghogho, 2022), evidence-based dataset (Sarrouti et al., 2021), annotation scheme (Bonet-Jover et al., 2021), automatically perform explainable fact-checking (Vedula & Parthasarathy, 2021), factoring fact-checks (Jiang et al., 2020), frame factuality assessment (Yao et al., 2021), ranked mistakes (Karagiannis et al., 2020), stance detection for fact checking (Shaar et al. 2021), cross-modal relations of entities (Springstein et al., 2021), extraction/clustering/contextualization (Smeros et al., 2021), autonomously discerning about the veracity of a given statement (Simpson & Yang, 2022), ClaimReview (Woloszyn et al., 2021), Sentence Retrieval task (Azevedo et al., 2022), Wikipedia knowledge base (Mykola & Saez-Trumper, 2021).

“Spotting” were mostly about 5G wireless network and labelled dataset (Pogorelov et al., 2021), BERT vector performing a Bag-Of-Words classifier (Sharma & Garg, 2021), differentiating false content from reliable news (Harrag and Djahli, 2021), Convolutional Neuron Networks (Alhindi et al., 2021), Fusion of syntactic features with BERT embeddings (Cheema, Hakimov, and Ewerth 2020)), embedding-based classifiers and BERT-based transfer learning (Wührl & Klinger, 2021), BERT variants as point-wise rerankers.(Mansour et al., 2022), DEI - debunking effectiveness index (Wang et al, 2021), dataset for evidence-based fact-checking (Sarrouti et al., 2021), fact-check labels for claims (Kotony & Toni, 2020), Leveraging in natural language processing(Sawczyn et al., 2021), XLM-RoBERTa Transformer architecture (Martín et al., 2022), public health table-based dataset (Carpuat & Meza-Ruiz, I.(2022), knowledge enhanced fact checking system (Zhu et al., 2021), Models based on BERT, LSTMs, and CNNs (Shaar et al, 2020), and detecting check-worthy tweets (Nikolov et al. 2020). The “monitoring” category has presented predictions using counterfactual explanations (Dai at. al., 2020); classification models (Casillas et al., 2022); expert annotators Nabožny et al., 2021); BERT to identify tweets (Moffitt, 2021); algorithmic fairness audits (Neumann et al., 2022); EQA - Extractive Question Answering (Rakotoson et al., 2022); MetaDetector (Ding et al., 2022); automatically labeled mappings (Srba et al. 2022); cross-

corpora abusive language detection (Shaar et al. 2021); interaction networks (Chen & Freire, 2020); automatic approaches to images on WhatsApp (Reis & Benevenuto, 2021); flagging Twitter bots (Lanius et al., 2021); supervised machine learning approaches (Simpson & Yang, 2022); text-based analysis (Lahoti & Prasad, 2022); and in-depth study of the number of retweets (Villar-Rodríguez et al., 2022).

For the category "publication" we were able to notice only six publications, which represented 9% of the total (

Figure **21**). This verification brought us a relevant statistic that shows us an area to seek a better understanding of why it is less explored and potential contributions that we can make in our research. Vedula & Parthasarathy (2021) introduce FACE-KEG, an application that automatically perform explainable fact checking. They present an output fact or claim, a model that constructs a knowledge graph. Gruz and Mai (2020) present a case study of a hashtag to propagate a false conspiracy theory and through automation verify the effectiveness of this case's disclosure content as the fake being more efficient than fact-checking publishing content. Babakar & Moy (2021) brings a theoretical review and FullFact's reports about the four phases in which fact-checking operates, in which it includes the "publication" phase as one of the highlights for automation. Ng & Carley (2021) analyzed fact-checked publications from Poynter and PolitiFact. Further, she broke them down in clusters by proposing a unique automated method, which can be used to classify diverse story sources in both fact-checked stories and tweets. We classified in this "publish" category the following two articles that described tagging fact-checking results with signaling, even though this is not text automation effectively featuring NLP or NLG techniques. Lanius et al. (2021) present how, under the COVID-19 infodemic, a flagging could be built into automated fact-checking systems and other misinformation abatement strategies. Likewise, Moravec et al. (2020) also present a fake-news flag solution in two theoretical routes presenting fact-checking results by automatic cognition and deliberate cognition.

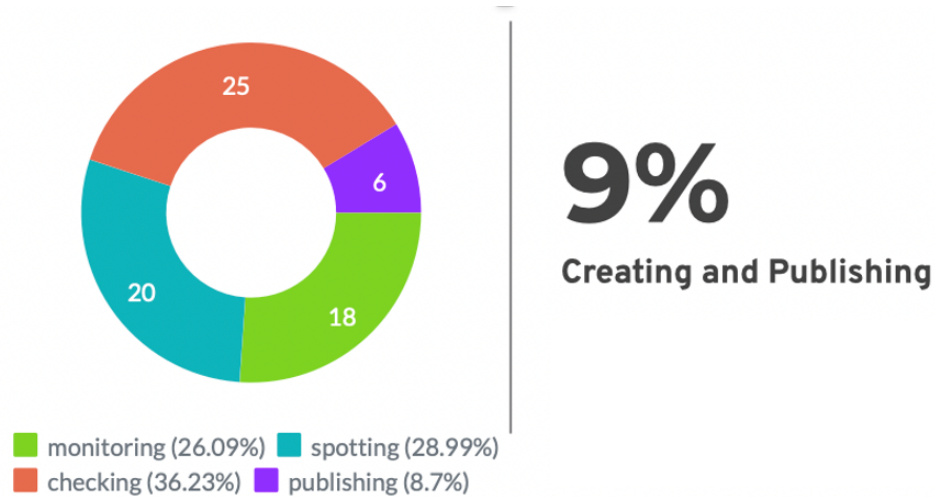


Figure 21. Less studied reported techniques according to the review

The TreeMap in Figure 22 highlights the combination of possible keywords, representing health fact-checking and technologies: Joining the term “covid-9”, “detection” and “misinformation” were the term most frequently used, followed by the term’s "claims" and "fact-checking". While the first corresponds literally to the word used in the research, the words fake, media, news, social, twitter, automatic, checking, disinformation, international, pandemic, conspiracy suggest technologies and themes addressed in the research publications.



Figure 22. TreeMap with keywords fact-checking during the pandemic period

The same was done with terms and their correlation (Figure 23) confirming that the most quoted words can be grouped in characteristics related to the pandemic period and misinformation, and a second group formed by words referring to congresses and procedures. The aim of the co-word analysis is to map the conceptual structure of a framework occurrences in a bibliographic collection. The clustering is to identify documents that express common concepts and are plotted on a two-dimensional map that, in our review, did relate the health fact-checking in one cluster and the academic fields in another. Based on the previous literature review, about automatic fact-checking, we could note that the “publishing” is the phase as an evolving area and has great scope to explore further.

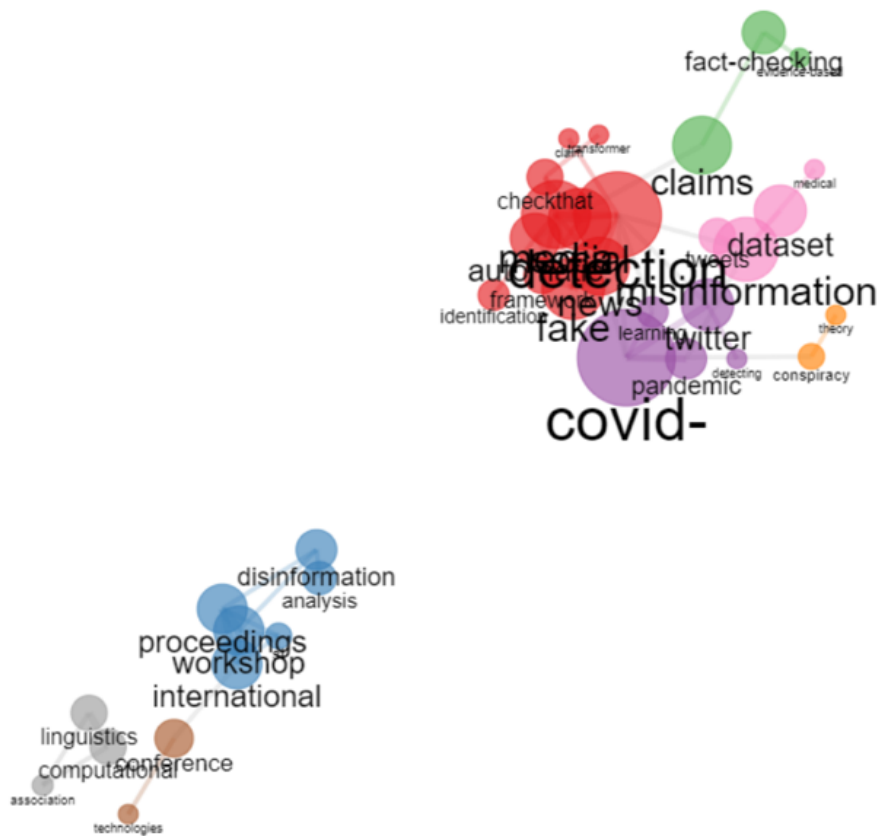


Figure 23. Keyword co-occurrences clusters with automatic fact-checking during the pandemic

We will therefore explore this together with AI in our qualitative triangulation in the next chapter. The research will present the methodologies that sought a performance possibility in the use of generative media to publicize fact-checking in the health area. This approach aims to use the large language model as a complement to traditional fact-checking work, like that of fact-checking agencies,

and to refer to the archives of a fact-check work with human supervision, to gather reliable information and generate content with less susceptibility to data selection bias or hallucinations, as described in the previous chapter as a current possibility for these models.

Chapter Summary

This chapter provided a comprehensive literature review on the intersection of digital ecosystems, health disinformation, and fact-checking. Initially, we examined the new digital ecosystem and emerging technologies, highlighting their influence on the post-truth era and information mediation. The discussion then moved to the challenges the misinfodemic poses and its impact on public health. We underscored fact-checking as a crucial tool for ensuring accountability in journalism and explored various AI-powered tools and their applications in combating disinformation. The chapter also addressed the ethical issues surrounding artificial intelligence, particularly in health communication, and the role of automation and generative media. This literature review concluded with a systematic review that delved deeper into the subject, identifying the area of publication as one that requires further study. This served as a guide for future phases to identify effective solutions to complement the existing body of knowledge.

Technology is the result of human imagination — of human beings envisioning alternatives to the status quo and acting upon the environment with the materials at hand to change the conditions of human and non human life.

(Friedman and Hendry 2019)

CHAPTER 3: Research Methodology

For this study, we will apply a qualitative methodology using a Value Sensitive Design approach. This exploratory study aims to familiarize ourselves with a problem to be analyzed further. Authors like Haguette (2013) emphasize that experimental research seeks a new perception of a specific phenomenon, making this approach suitable for fostering an innovative tool. Qualitative procedures differ from quantitative research methods in their philosophical assumptions, data collection, analysis, and interpretation methods. Although both processes share similarities, qualitative procedures, which rely on image and text data, have unique steps in data analysis and draw on diverse inquiry strategies (Creswell and Creswell 2017).

The qualitative inquiry strategies chosen are based on the landscape of systematic procedural guidelines (Strauss and Corbin 1997). All perspectives vie for center stage in this unfolding model of inquiry known as qualitative research. We will use interpretative research, focusing on the experience with participants by the inquirer, which introduces strategic, ethical, and personal issues into the qualitative research process (Locke, Feldman, and Golden-Biddle 2015).

According to Creswell and Creswell (2017), sequential mixed methods procedures, such as those used here, can be followed by a researcher who seeks to elaborate on or expand the findings of one method with another. This may involve starting with a quantitative approach where a theory or concept is tested, followed by a qualitative approach involving a detailed exploration of a few cases or

individuals. This chapter begins by revisiting some key concepts of the Value Sensitive Design methodology. It then introduces the framework devised to adapt the tripartite approach to our case study and the issues it addresses. Our innovation consists of redefining this methodology within the context of using fact-checking and inoculation as preemptive measures to be tested. Each phase of this experimentation will be detailed in the following sections.

Section 3. Methodological approaches in a Value Sensitive Design study

We will apply the Value Sensitive Design (VSD) methodology as a foundational framework to guide our investigation and design processes. As introduced in the first chapter, VSD is an approach that integrates human values directly into the core of technological design, asserting that these values are intrinsically linked to the socio-cultural contexts from which they emerge (Umbrello 2021). This methodology recognizes the diversity of human values and places them within the broader context of societal and technological systems. By applying VSD, this research aims to navigate the complex interplay between technology, society, and human values, ensuring that the outcomes are both ethically informed and culturally relevant.

VSD is structured around six core principles: the interactional view of technology and society, a tripartite methodology encompassing conceptual, empirical, and technical investigations (Figure 24), the inclusion of diverse stakeholders, the navigation of value tensions, considerations for multi-lifespan design, and a focus on progress over perfection (Friedman and Hendry 2019). Each of these principles will play a critical role in shaping our technology design proposal in response to human needs, the societal challenges of the infodemic, and technological advancements with AI.

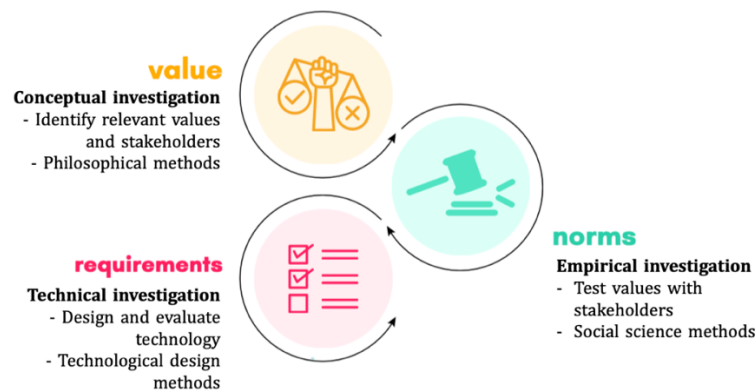


Figure 24. VSD tripartite approach

By following the VSD approach, this research endeavors to produce a proposal solution that is innovative, practical, ethically grounded, and socially responsible. The following sections will delve into the application of VSD within our research, highlighting how each of the methodology's foundational principles will inform and guide our investigation. The Value Sensitive Design (VSD) methodology actively encourages practitioners to engage with its theoretical underpinnings, promoting a continuous cycle of improvement through practical application. Far from being a rigid set of tools, VSD is meant to be adapted and woven into the fabric of existing design practices, thus enhancing its widespread applicability (Friedman and Hendry, 2019).

Historically, VSD scholars have introduced a wide array of methods and tools to support the approach, including stakeholder analysis, value scenarios, ethnographically informed inquiries into values and technology, value-oriented semi-structured interviews, value-oriented coding manuals, multi-lifespan co-design or prototyping, envisioning cards, and diverse voices, among others ("VSD Lab" 2020). In this thesis, we propose to integrate a selection of these VSD methods with traditional qualitative methodologies. Our approach will be outlined across three stages, adhering to VSD's tripartite framework: conceptual, technical, and empirical investigations. This structure will guide our exploration of how these methods can be effectively applied within the context of our research.

For the investigation to be carried out and to obtain the necessary results applied to the present reality, we specified that it would be relevant to analyze the

tools that have already been created within the journalism scope (Mrowczynski 2013). This logic will be essential to draw an inventory of valuable tools for fact-checking and synthetic media. In this context, Diakopoulos (2016) is highlighted for his analysis of platforms and devices in the media scope.

The first process associated with this methodology is identifying the problem to be studied. We decided that the first research question considers the field of fact-checking and the difficulties of validating the truth and making it accessible to the public. To validate and iterate solutions that represent the real needs of communicative health issues, we conducted a systematic review on the theme. This review aims to consider the nuances involved in the complex issue of health misinformation and the emergence of fact-checking as a mitigation strategy.

Faced with a challenging task, the choice of this topic aims to provide a panoramic view of the theme amid the challenges posed by health journalism and AI ethics. We narrowed down the themes by delving into some ethics methodologies and developing a more careful analysis of the challenge in automatic fact-checking through the PRISMA methodology (Moher et al. 2009b), offering a significant and original contribution to the problem.

Based on the interviews and case studies process, the second step of the methodology will analyze fact-checking and synthetic media to answer the second research question. Using case studies and interviews in research is crucial for gaining in-depth insights and understanding complex phenomena in real-life contexts (Mojtahed et al. 2014). Case studies allow researchers to explore a specific subject or situation in detail, capturing the richness and intricacies of the topic under investigation (Hancock, Algozzine, and Lim 2021).

Through interviews, researchers can directly engage with participants, obtaining first-hand perspectives and experiences, and providing valuable qualitative data that complements quantitative data (Ge and Land 2003). This combined approach offers a holistic view of the research problem, generating nuanced and contextually relevant findings that can lead to the discovery of new patterns, themes, and underlying factors. Knowing cases will help us implement fact-checking solutions targeting issues on health data, such as vaccination and Covid-19.

The third step will apply methods from the VSD theory along with focus groups and surveys. The approach aims to develop and integrate ethical considerations, user values, and stakeholder perspectives into the design of technological solutions, products, or systems.

Thus, to achieve the objectives proposed here, only the activities related to the tasks are contemplated: (a) theoretical research on the scientific relevance of information; (b) delimitation of the characteristics essential for the definition of the basic technology of these items; (c) selection of the most appropriate applications for the use of health fact-checking communication; (d) description of their primordial meanings and their extensions, with the selection of examples; and (e) identification of the scientific validity of the data presented. To answer the fourth research question, we will propose the iteration and confirmation of an experience testing a non-functional prototype to validate the standard workflow of fact-checking and alerts. This phase considers inviting previously identified stakeholders to conduct tests with the specified target group to test the solution.

Since misinformation in health has an intense connection to social relations and public interaction, whether digital users appreciate avatars and automatic content remains to be seen. Depending on its acceptance and applicability as a tool to help citizens, the final considerations required to conclude this study will be determined as an effective plan for the future media scope. At the end of these experiments and interactions, it will also be required to review the entire procedure of organizing investigative data from a theoretical point of view. In this path, it will be possible to evaluate the improvements made in the state of the art and how the processes involved in the system's design contributed to the development of scientific knowledge, particularly in automated journalism and health fact-checking. Next, we detail the methods used for the conceptual approach.

3.1 Phase 1 – Conceptual Investigations

This phase is designed to scrutinize and uncover new possibilities through the use of fact-checking to communicate health risks effectively. Conducting a systematic review (Kitchenham and Charters 2007) is crucial as it offers a rigorous

and comprehensive synthesis of existing research on a specific topic. This method enables researchers to identify patterns, trends, and gaps in the literature. By systematically collecting and analyzing all available evidence, a systematic review minimizes bias and strengthens the reliability of the conclusions. It provides a solid foundation for evidence-based decision-making and policy formulation.

3.1.1 Applying the PRISMA Protocol to Narrow the Thematic

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol will be employed as its methodology is essential in research and evidence synthesis. It offers a structured and transparent approach to conducting systematic reviews and meta-analyses, ensuring studies are performed rigorously, consistently, and clearly (Moher et al. 2009b). Following the PRISMA guidelines helps researchers reduce bias and enhance the reproducibility of their findings, thereby improving the overall reliability of the research process. This methodology advocates for a comprehensive and systematic search for relevant studies, aiming to capture all available evidence on a specific topic.

Moreover, deductive science (Hall, Savas-Hall, and Shaw 2022) plays a crucial role, as it employs a structured and logical approach to hypothesis testing and theory development. Starting with general principles or theories, it tests specific hypotheses derived from them. Deductive reasoning enables researchers to establish causal relationships, predict outcomes, and confirm or refute theoretical propositions. This robust method increases the credibility and validity of scientific findings, leading to a deeper understanding of natural phenomena and fostering advancements in science and technology.

Together, systematic reviews and deductive science are pivotal to rigorous scientific inquiry, significantly contributing to the progress and reliability of knowledge across various academic disciplines. Integrating these elements into the literature review will provide a grounded and pertinent state of the art, addressing a complex yet critical topic. Despite the challenges of amalgamating technology theory, ethics, and disinformation, it remains a vital endeavor.

3.2 Phase 2 – Technical Investigations

As a second phase, we now focus on investigating technological concepts by analysing related cases and interviewing stakeholders. we will delve into specific cases and their correlations with fact-checking and generative media. Case studies (Kerlinger 1979) will be supplemented by interviews (Cervo, Bervian, and Da Silva 2009), focusing on qualitative strategies to understand the central phenomenon of customer relationships within fact-checking and automation. We have observed some engaging real experiences related to the subject, leading us to adopt a qualitative approach that explores the relationship of AI synthetic media with user experience and perceptions of accuracy.

3.2.1 Case Studies Using Automation and AI

This study aims to understand how citizens relate to automatic fact-checking and its digital biases, particularly regarding its contradictory points. Recognizing digital limitations is crucial for achieving a comprehensive understanding of current approaches in the humanities. Data triangulation, collected from various sources including interviews and case studies, will validate the research (D.T. Campbell and Fiske 1959). Following Haguette's (2013) guidelines, a case study on the chatbot designed for fact-checking news and the newscast created using deep learning and 3D characters will be conducted, addressing the third Research Question: "How effective is the chatbot in fact-checking news and the newscast created using deep learning and 3D characters?". This involves determining the context, developing data collection instruments, collecting data, applying content analysis to extract meaningful insights, and concluding the effectiveness of the chatbot and newscast based on data interpretation.

Content analysis within cases will be applied to validate observations from other analysis methods, observe the recurrence of significant elements related to digital and new services, and cross-reference data from reports and texts with observations of techniques used and workshops conducted.

The discussion plan for data analysis may include multiple components, involving the process of making sense of text and image data, preparing data for

analysis, conducting in-depth studies, and interpreting the data's broader meaning. This process is likened to peeling back the layers of an onion by some qualitative researchers, representing the data, and interpreting its significant meaning. Various generic processes outlined in the proposal convey an overview of the qualitative data analysis activities (Rossman and Rallis 2010).

3.2.2 In-depth Interviews and the World of Automation

Value Sensitive Design (VSD) recommends employing semi-structured interview questions to deeply explore stakeholders' perceptions, opinions, and values regarding a technology (Borning et al. 2005; Friedman 1997). This method is particularly highlighted when it's essential to understand stakeholders' nuanced evaluative judgments about a technology and their underlying reasoning (Friedman et al. 2006). VSD also emphasizes the importance of exploring any additional insights or considerations from stakeholders, ensuring a comprehensive understanding of their perspective on the technology (Czeskis et al. 2010; Freier 2008).

Considering the systematic review will raise study topics and case analysis will list solutions for the proposed topic, interviews will seek more precise elements about challenges and resolutions. Given the delicate and complex nature of this thesis topic, questions will delve into deeper issues, aiming to avoid personal value biases. Interviewer bias, referring to the impact an interviewer's personality may have on respondents' answers (David and Sutton 2004), will be carefully managed.

This method does not aim to uncover a single answer but to gather validations and quality opinions on a problem, addressing the paradoxical connections between technology and journalism, and the sciences (Kvale 1996). Qualitative interviews may involve conducting face-to-face interviews with participants, telephone interviews, or engaging in focus group interviews with at least six participants (Kvale, 1996), focusing on open-ended questions designed to elicit participants' views and opinions. This work will utilize the problem-centered interview, based on a procedural model, to direct questions towards understanding the facts and processes (Mrowczynski 2013). During the study, qualitative responses may be collected in a semi-structured interview format, guided by a list

of questions or topics (Kvale, 1996). Below we explain how these technological approaches were finally scrutinised with an empirical eye.

3.3 Phase 3 - Empirical Investigations

In this phase, we will organize collaborative workshops where stakeholders co-design ideas, prototypes, and solutions. These workshops promote open dialogue, foster empathy, and facilitate shared ownership of the design process. Co-design workshops are collaborative gatherings where stakeholders, users, designers, and other relevant parties share their perspectives, expertise, and insights. Participants contribute to the co-design process, helping to shape the technology in alignment with their values and concerns.

The concept of co-design, also known as co-creation, originated in the late 20th century, evolving from the foundations of participatory design (PD) and user-centered design (UCD). Defined broadly, co-design encompasses collaborative creative efforts where both professional designers and non-professional end-users, along with potentially other stakeholders, engage collectively in the design process (Sanders and Stappers 2008). Over recent years, the acceptance of co-design as a crucial methodology for user studies has grown across various domains, including product and service design, interaction design, and HCI - human-computer interaction (Muller and Druin 2012).

Building on this framework, Sanders and associates have further developed the notion of co-design spaces, emphasizing the significance of structuring these environments to enhance the creative output of participants with limited design experience (Sanders and Westerlund 2011). This structure involves the provision of resources, guidelines, and activities within the co-design setting to promote the exploration of diverse solutions, encourage the sharing of ideas, and foster the pursuit of innovative concepts. Additionally, Sanders and team illustrate that established methods like “Envisioning Co-design Workshops” and specific design tools for creative engagement can effectively serve as embodiments of co-design spaces (Sanders and Stappers 2008).

This research phase proposes data compilation using some specific VSD methods (Friedman, Hendry, and Borning 2017) along with focus groups, surveys, and co-design methods. Combining the strengths of these methodologies can lead to a more comprehensive and human-centric design process. This introduction outlines a method inspired by value-sensitive design (R. Pereira and Baranauskas 2014), integrating the co-design approach, technological probes, prototyping, and evaluation.

In exploratory research (Swedberg, 2020), and particularly within these workshops, our aim is to engage stakeholders involved in fact-checking and to validate their perspectives on automation, drawing upon exploratory and descriptive Value-Sensitive Design research (Cenci and Cawthorne 2020). A critical component of this phase involves evaluating solutions through technological probes and non-functional prototypes. This facet of the research delves into technology probes and prototypes as methods in technology design and user engagement, informed by the foundational work of Hutchinson et al. (2003) and Güldenpfennig et al. (2016) on technology probes, and insights from Clement (1996), Muller (1999), and Gaver, Dunne, and Pacenti (1999) regarding prototypes. Both technology probes and prototypes are integral to the design process, albeit serving differing purposes and exhibiting unique characteristics.

Technology probes involve installing technology into a real-world context, where researchers observe its use over time. The primary objective is to gather information about the users' interactions with the technology and to inspire ideas for future technological developments (Hutchinson et al., 2003). Unlike prototypes, technology probes are technically simple and flexible regarding their possible use. They are not fully developed products but tools to explore and determine which types of technologies might be of interest for future design (Güldenpfennig et al., 2016). According to Gaver, Dunne, and Pacenti (1999), this could be highlighted:

The probes were part of a strategy of pursuing experimental design responsively. They address a common dilemma in developing projects for unfamiliar groups. Understanding the local cultures was necessary so our designs wouldn't seem irrelevant or arrogant. Still, we didn't want the groups to constrain our designs unduly by focusing on their already-understood needs or desires (Gaver, Dunne, and Pacenti 1999, 21).

The simplicity of technology probes is deliberate, often limited to a single primary function and few accessible features. This contrasts with prototypes, which typically encompass multiple functionalities to meet a wider array of needs, although not all functionalities may be fully developed (Clement 1996). Whereas prototypes are iteratively refined based on user feedback, technology probes remain static during their deployment, sometimes incorporating intentional limitations to elicit user responses (Güldenpfennig et al. 2016). The capability of technology probes to log user interactions facilitates the generation of new ideas and fosters discussions between users and designers, an aspect not primarily focused on by prototypes (Hutchinson et al. 2003).

Our methodology will proceed through the following steps: a) Identifying stakeholders relevant to the technology's development and potential impact; b) Co-Design Preparation, which includes inviting stakeholders to participate and ensuring diversity and inclusivity among workshop participants to represent a broad spectrum of perspectives; c) Value Elicitation, using various methods to surface and prioritize participant values, such as group discussions and brainstorming sessions; d) promoting exercises that encourage creativity and the empathetic understanding of user needs; and e) Iterative Co-Design, where participants collaboratively refine and evolve design concepts based on identified values and insights. Finally, f) Value Integration and Evaluation involves incorporating the values identified during the workshops into the technology's design.

In this empirical phase, it is crucial to implement user testing strategies thoughtfully. The introduction of novel interface designs can present substantial usability challenges. Unlike conventional designs, these new interfaces might not align with users' pre-existing knowledge, potentially leading to an increased occurrence of usability issues (Ritchie Macefield 2006; Raskin 1994; Nielsen and Landauer 1993). Early prototypes featuring novel concepts may, therefore, necessitate fewer participants to identify significant usability problems, as the degree of novelty inversely correlates with the number of required participants.

Research by Nielsen and Landauer (1993) suggests a minimum of at least eight participants to achieve significant findings. Studies with around 25 participants are likely to yield statistically significant results, with 10-12 participants often being a

good number for such innovative findings, aligning with recommendations from (Spyridakis and Fisher 1992) Therefore, when seeking significant outcomes, selecting a study group size within the 8-25 participant range, ideally around 10-12, is advisable based on the collective guidance from these studies (Ritch Macefield 2009). These considerations will be considered in our research.

In the upcoming presentation of our research, we will detail how we have seamlessly integrated three methods into a co-design environment. This approach is aimed at stimulating stakeholders to thoroughly consider the technical attributes and value propositions of their design initiatives.

3.3.1 Value Scenarios

The concept of Value Scenarios is a method advocated by Value Sensitive Design (VSD) as an augmentation of the conventional scenarios of use and problem-solving (Carrol 1999). These narratives are crafted to highlight the societal dimensions of emerging technologies, with a particular focus on their long-term repercussions and potential adverse effects (Nathan et al. 2008). In earlier studies, value scenarios have been utilized by designers as a tool to investigate values around responsible parenting and the use of mobile apps for supervising teenagers (Czeskis et al. 2010). Conversely, in some cases, stakeholders themselves have authored value scenarios, employing them as a means to gather narratives that spotlight safety concerns among homeless youth regarding mobile phone usage (Woelfer et al. 2011).

These narratives reveal the human and technical facets of technology within its context, aiming to shed light on the consequences for both direct and indirect stakeholders, underscore associated core values, and explore the broader, systemic implications of widespread technology use over time. In our refined approach, we plan to examine technological fact-checking probes through a broad stakeholder lens, including researchers and specialists. The goals of our research include:

- a. Review of Existing Probes: Participants will critically review current technological probes to discern their ethical ramifications, alignment with values, and opportunities for enhancement.
- b. Methodology Testing: By examining selected probes, participants will unearth potential benefits and drawbacks. They will evaluate the methodology's capability in revealing ethical dilemmas and clashes in values, suggesting improvements as needed.
- c. Envisioning Cards: Utilizing envisioning cards— a set of 28 themed cards designed to prompt consideration of long-term and systemic technological impacts—the participants will be encouraged to think critically about the broader consequences of technology (Nathan et al. 2008).
- d. Technological Requirements: Drawing from the elicited values and norms, participants will derive specific technological requirements for the probes, ensuring they reflect the identified ethical standards.

3.3.2 Focus Group Discussion

Given the profound connection between health misinformation and its impact on social relations and public interaction, it's pivotal to assess digital users' reception towards avatars and automatic content within a group dialogue framework. Focus groups, as defined by Krueger (2014), are structured group debates designed to probe a series of issues, where the researcher engages with at least six participants in face-to-face discussions. These dialogues typically feature unstructured and open-ended questions crafted to draw out the participants' perspectives and opinions (Barker and Rich 1992). In this research, we will adopt a problem-centered, value-oriented debate approach, informed by a procedural model, to develop theories on how questions are directed towards understanding the design and communication outcomes associated with synthetic media (R. Mrowczynski, 2013).

The rationale behind utilizing focus groups stems from their ability to deeply explore the qualitative dimensions of a specific social group's viewpoints. Such settings allow for an in-depth examination where numerical representation is

secondary to the richness of participants' shared experiences, discussions, and interactions . Departing from quantitative metrics and empirical proofs, focus groups offer a wealth of qualitative data, drawing on a variety of approaches to unearth insights (M.d.A. Marconi and Lakatos 2001). The upcoming segments will address these critical aspects:

- a. **Fit Assessment Techniques:** The focus group will introduce and explore various fit assessment techniques, such as Heuristic evaluation, Usability testing, Ethnographic methods, and Experience sampling.
- b. **Normative Framework:** Leveraging the values brought to light, participants will collaboratively forge a normative framework to guide the ethical dimensions of the design process.
- c. **Iterative Validation:** The established values, norms, and technological requirements will be subject to ongoing validation and refinement, enriched by diverse perspectives through continuous dialogue and feedback.

3.3.3 Value-oriented Mock-up, Prototype and Field Deployment

This method is recommended by the VSD lab's directives ("VSD Lab" 2020) integrating identified values into the development of a proposal mock-up, leading to the formulation of normative guidelines and recommendations. According to Freier (2008), the development, analysis, and co-design of mock-ups, prototypes, and field deployments are instrumental in scaffolding the exploration of value implications for technologies that are either in the conceptual stage or not yet widely adopted (Yoo et al. 2013). These mock-ups, prototypes, or field deployments are designed to highlight the potential impacts on both direct and indirect stakeholders, elucidate value tensions, and position technology within human-centric contexts.

This phase of the methodology involves co-design and empathy building (Dimitrakopoulou and Lewis 2023). The process fosters empathy and enables the identification of values and ethical considerations. Again in this, technological probes and ideation techniques can be employed to explore innovative solutions

(Portugal et al. 2023). Technological probes are the experimental prototypes or technologies that prompt user feedback and generate insights. This approach encourages stakeholders to envision possibilities and creatively brainstorm ideas that align with the values uncovered in the first phase.

The stakeholders focus on evaluating the non-functional prototypes with a keen eye on the identified values and ethical implications. Feedback is collected from stakeholders to assess the effectiveness and alignment with the initial empathetic understanding. Based on the evaluation results, the design is iteratively improved to better address the users' needs and values. By the end of this methodology process we will have a solution mock-up inspired by values, norms, and technological validation. In the last process we will also have the output with ethical remarks as guideline for the proposed AI health disinformation fact-checker solution. This method will involve the following activities:

- a. **Questionnaire Evaluation:** Employing qualitative behavioral user testing (Nielsen 1994), participants interact with various chatbot prototypes, providing detailed feedback on values and effectiveness. Questionnaire drawn up with 13 questions in line with Kitchenham and Pfleeger's (2002) recommendations, to be explained further in the next part of the research. The responses are on a Likert scale, likely ranging from 1 (do not agree at all) to 5 (completely agree). Participants can rate which proposed prototype solution is their favorite.
- b. **Brainstorming and Empathy Map:** Participants engaged in co-design activities to collaboratively define the persona and characteristics of the chatbot avatar. The Empathy Map (Siricharoen 2021) entails developing a visual map to consolidate insights from user into sections like seeing, hearing, and feeling, facilitating the creation of detailed personas. These personas offer a comprehensive understanding of user motivations and challenges, guiding user-centered decision-making in the design process.
- c. **Ethical Implications:** Ethical considerations related to the evaluation process will be thoroughly discussed, and guidelines for ethical evaluation will be formulated.

This proposed methodology utilizing Value-Sensitive Design principles and the selected methods provides a structured and systematic approach to the ethical

development and evaluation of technological probes. By incorporating diverse perspectives and stakeholder involvement, this approach aims to produce responsible and value-driven technological solutions aligned with societal needs and aspirations. The iterative and collaborative nature of the workshops will foster a deeper understanding of value implications and guide the creation of more responsible and acceptable technological solutions.

As for the data collected in the questionnaire, despite the small number of ranges covered, we carried out Exploratory Data Analysis (Tukey 1977) to obtain information from the data. As stated above, our original intention was not to test the usability and speed of the application, but to create a ranking of the most attractive solutions and try to find different patterns and relationships in the data using some statistical graphs and other visualization techniques.

The outline of the interviews, workshop testimonies, and focus groups will be evaluated using reflexive thematic analysis (Braun and Clarke 2014; Clarke and Braun 2013). According to Clarke and Braun Clarke and Braun (2017, 67), “typically start with theme development following some data familiarization or with the themes already conceptualized prior to data analysis (e.g., the data collection questions are used as the themes). So although themes are the output of analysis, they are also an input into the analysis that delimits the scope and focus of the analysis”.

Considering the use of Value Sensitive Design (VSD) and an initial phase to identify fundamental values to be incorporated through participatory means, our research will primarily adopt an inductive approach. As Clarke and Braun note, “Being purely inductive is impossible, as we always bring something to the data when we analyze them” (Clarke & Braun, 2017, p.69). Despite this, the inductive method is chosen because we believe in starting from a bottom-up approach driven by participatory data. Given the innovative nature of our approach, this method is more suitable for exploratory research, which is the focus of our study (Clarke & Braun, 2017).

In this research, we will employ a reflexive method inspired by Clarke and Braun (2017), consisting of six phases: (a) familiarization with the data, (b) data coding, (c) initial theme generation, (d) review and development of subthemes, (e)

refinement and renaming of themes, and (f) production of the final report with tables summarizing the dataset's findings. The overarching themes will be drawn from our research questions and will be used to identify our dataset using NVivo 14 to guide the answers to the research questions through the transcripts, grouping them into themes appropriate for the Value Sensitive Design methodology, subdividing or merging them whenever the grouping unit is too broad or too narrow. The four fundamental themes selected from the research questions are: Health Misinformation, AI Ethics, Fact-Checking, and Artificial Intelligence. To this end, qualitative interviews will be conducted based on a semi-structured questionnaire with questions organized into "Basic Descriptive," "Experiences," "Clarification," and "Comparison." The final column provides examples of the interview script, that will be supplemented with elements introduced by the analysis of the empirical material (Table 5).

Table 5. Code framework with examples of the interview script

	Codes	Themes	Questions	
Research Question # 1	Health Misinformation	Effects on public health decisions and behaviors	RQ: How can integrating artificial intelligence in disseminating fact-checking outcomes contribute to mitigating misinformation related to health on digital platforms?	
			Types of Questions	Clarification Could you please describe the reasons for applying ethical implications here?
				Comparison Do you have any comments on the tensions and advantages of this ethical pervasiveness?
				Experiences Can you give me an example of perceptions of ethical issues?
Research Question # 2	Artificial Intelligence	Contributions of automation to productivity increases Ensuring fairness in AI applications Technology advantages	RQ: How can artificial intelligence facilitate disseminating authenticated health information?	
			Types of Questions	Clarification Could you please explain more about what you you like in this solution.
				Comparison Do you have any comments on the differences and advantages between the solutions?
				Experiences Can you give me an example of perceptions of what you think that is relevant about the health disinformation?
Research Question # 3	Fact-checking	Development and effectiveness	RQ: In what ways can the process of health fact-checking benefit from the incorporation of artificial intelligence to speed up the dissemination of verified information on digital platforms?	

		of fact-checking tools		Basic Descriptive	Would you like to give details about the solution that you would like to describe as more relevant?
			Types of Questions	Clarification	Could you please describe the essential elements for your chosen solution?
				Comparison	Could you compare the essential values to be addressed?
				Experiences	Can you give me an example of perceptions of using your solution?
Research Question # 4	Ai-Ethics	Participatory Governance	RQ:How can integrating artificial intelligence promote ethically responsible practices in health fact-checking, explicitly focusing on disseminating counter-information?		
			Types of Questions	Basic Descriptive	Would you like to start by telling me details about this Project, such as the time spent on each activity?
				Clarification	Could you please describe the methods for developing?
				Comparison	Do you have any comments on the differences and advantages of this process to a similar one done manually without AI?
				Experiences	Can you give me an example of perceptions of using your solution?

These methodological phases will be applied to a group of stakeholders that will be explained in the next section. Understanding this group and the rationale behind it is important in our thesis, as the methodologies themselves consider it important to understand direct and indirect stakeholders. But mainly because this is a complex problem and narrowing its scope is fundamental to giving valid answers to the problem.

3.4 The stakeholders involved

The concept of directly and indirectly involved stakeholders is crucial in the framework of VSD. Friedman et al. (2006) highlights the importance of acknowledging not only those individuals who interact directly with a technology (for example, someone in an office benefiting from a webcam overlooking a local) but also those whose data or presence might be affected by the technology's deployment (such as the privacy and security concerns of individuals captured by

the webcam). As part of the methodology, we need to clarify which groups/entities/citizens from the health infodemic ecosystem will contribute to the interviews and participate in the workshops.

These stakeholders are divided into four groups (Figure 25) and will be described below. It should be noted that the stakeholders above can be promoters of behavioral change in a logic of health journalism against misinformation. This proposed intervention ecosystem is inspired by UNESCO's Disinfodemic study (Posetti and Bontcheva 2021), which describes these four categories as umbrellas for action to combat disinformation in the pandemic period: identifying, distributing, supporting and production. Our study used identical terms to rationalize the group of stakeholders to be consulted, with the fact-checking platform being thought of as the central tapering channel.

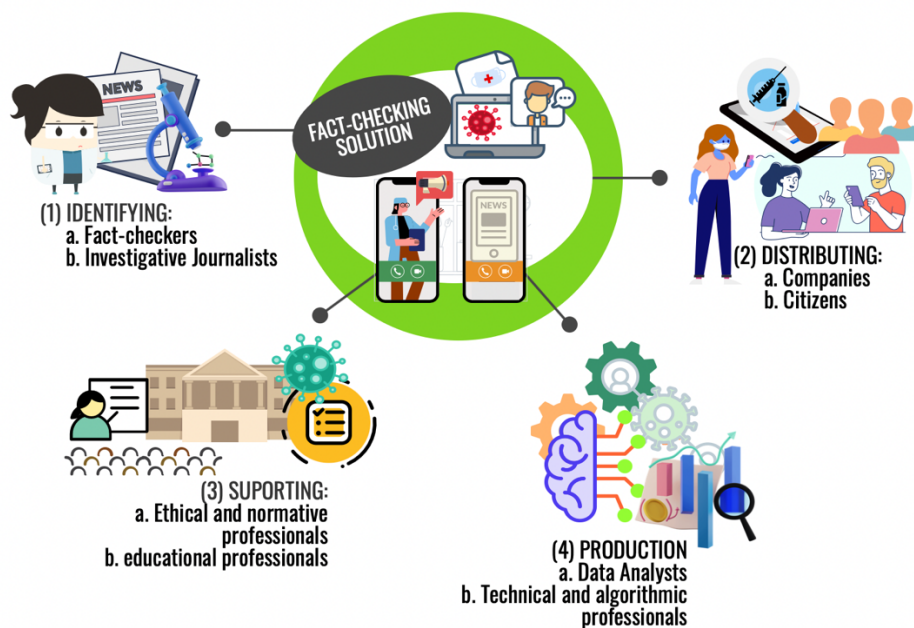


Figure 25. Four main groups of stakeholders to be included in the research

As explained, the value sensitive design methodology identifies finding stakeholders as an important phase (Friedman and Hendry 2019). Still, critics of the method (M. van Steenberg and van der Spoel 2021) point out that one of its problems is that it does not delimit a stakeholder boundary and can be a problem

in achieving good results. Therefore, we have already used an established study from the area of health disinformation to have these four groups well identified and explored in the phases of our research. Therefore, four stakeholders groups are considered essential in the process, although others can be understood as complex and transversal as we are treating.

IDENTIFYING is the first category (1) and incorporates fact-checkers, debunkers, and journalists as one category of this axis and, as central elements of the performance, they can validate information on health in different formats and for other new technological media. The DISTRIBUTING (2) category represents citizens and companies who receive fact-checked information and then decide whether or not to share it with others. The SUPPORTING (3) category relates to ethical and educational responses' empowerment and credibility label efforts. The PRODUCTION (4) category relates to generative media tools and the latest technologies surrounding their appearance. These groups come close to the embodiment of infodemic ecosystem and might be a starting point for validating values and framework for the proposed no-functional prototype. In this study, these concepts will be used as a way of seeking saturation, which, in qualitative research, refers to the point at which no new data or knowledge is generated from further data collection (Guest, Bunce, and Johnson 2006). Saturation is intended to be reached when the categories of identifiers, supporters, production and distribution - terms defined by the UNESCO study - have been fully explored and no additional relevant information has emerged.

3.4.1 Citizen's Technological Perceptions: EducAge and Elderly Caregivers

The "DISTRIBUTING" category pertains to what we identify as citizens and companies. From the perspective of citizens, we see this group as the primary beneficiaries of our solution, defined by VSD as direct stakeholders, aimed at addressing the impact of health misinformation. Our goal is for them to reap the potential benefits of the solution we plan to prototype. To select and justify our target audience more precisely, our research has been directed towards gathering data that supports our choice.

In recent years, studies have highlighted areas suffering from a lack of adequate news coverage, termed as news deserts (Abernathy 2023). These deserts represent a significant fallout from the crisis facing local journalism, particularly impacting smaller communities. In Portugal, the "News Deserts Europe - Portugal" report by Jerónimo, Ramos, and Torre (2022) notes that over 80% of such news deserts and semi-deserts are located in Portugal's North, Center, and Alentejo regions, with Beja, Bragança, Évora, Portalegre, and Vila Real experiencing the highest prevalence of these conditions. Recognizing the challenges these deserts pose to communities, our research aligns with the European EducAGE ("Educage Project" 2023) project aimed at addressing this issue in the region, particularly focusing on Bragança, Portugal. The EducAGE project is lead by an European consortium including six institutions from five committed countries to tackling key challenges in adult education, especially around Healthy Ageing and mitigating health misinformation among the elderly within a socio-economic framework.

The primary goal of EducAGE is to improve the skills and knowledge of educators and care providers in Healthy Ageing and elderly care. It plans to develop a specialized educational program to enhance the capabilities of caregivers and social workers, improving care quality for the elderly. Caregivers, who support individuals in daily living activities, range from family members to professionals, providing both formal (paid) and informal (unpaid) care. Informal caregivers, making up 80% of long-term care providers in Europe, often lack necessary training. The project adopts a modern approach, focusing on Healthy Ageing and empowering caregivers to support dignified ageing. It includes developing a curriculum based on WHO guidelines, employing mobile and e-learning platforms, and integrating digital tools for Healthy Ageing.

Recognizing digital literacy variances, part of the Educage program focus on enhancing digital skills among caregivers, particularly women over 65, promoting gender equality and digital competence among educators. The EducAGE project aims to offer comprehensive training to caregivers, fostered by collaboration among academic institutions, the private sector, municipalities, and associations. Their innovative course, designed to address the unique challenges of elderly care, is

made freely available, contributing significantly to adult education and caregiver support.

Although several European and national strategies have been made available, most caregiver training programs still lack digital readiness and the ability to integrate relevant digital content on healthy aging and tools into their daily activities (Furtado et al. 2022). Given these common objectives and understanding that our project has a solution that can serve the same purpose as the Educage project, we proposed carrying out our research with the public reached by them. Therefore, focusing on the DISTRIBUTING category, and receiving the agreement for this joint research, we carried out our tests with caregivers of elderly people who are assisted by this European project in the town of Macedo de Cavaleiros, in Bragança.

When it comes to companies, our attention is centered on those involved in communication and journalism, as well as entities operating within the digital communication and social media sectors. We will ensure the confidentiality of information shared by interviewees, and to foster a more open dialogue, individual names will not be disclosed in any publications. This approach to anonymity is designed to encourage interviewees to freely and fully discuss their experiences and the nature of their work within the industry, ensuring an environment of equity in expression (Sousa 2005). In light of this broad scenario, although we have concentrated our efforts on a specific aspect of health misinformation—namely the use of generative AI for fact-checking—it is essential to analyze the associated research risks and challenges in the next part.

Portugal stands out as a notable example regarding news trust and the interaction with illegitimate content. Despite a high level of trust in news — 62% of respondents trust news, comparable to Finland (Newman et al. 2018) — there is significant concern about misinformation. The Reuters Digital News Report 2018 indicates that 71.3% of Portuguese respondents are worried about the veracity of online content, with 76% having recently encountered dubious content. Trust levels differ greatly by source: while 48% trust news found through search engines, only 29% trust news from social media platforms. Additionally, only 8.6% of respondents paid for online news in the past year, and 31.1% use adblocking software (Paisana, Pinto-Martinho, and Cardoso 2020). The report also utilized a

news literacy framework by Maksl, Ashley, and Craft (2015), showing that higher news literacy is linked to more critical media consumption. Although Portugal's patterns align with those of other countries, there are fewer respondents with very high news literacy, suggesting an area needing attention (Paisana, Pinto-Martinho & Cardoso, 2020).

3.5 Methodological Challenges and Mitigation Strategies

In undertaking the integration of human values into the technological design of a health fact-checker verifier via the Value Sensitive Design (VSD) framework, we inevitably encounter a set of methodological challenges. A primary obstacle arises from the intricate process of identifying and aligning the various values held by different stakeholder groups. Given the potential for conflicts or competing priorities among these values, finding a path to their seamless integration into technology design is not straightforward. Our approach to mitigating this complexity involves an iterative engagement with stakeholders, leveraging methodologies such as value mapping and scenario development to foster consensus and prioritize universally valued principles.

Ensuring comprehensive stakeholder representation also stands as a challenge. Particularly for those groups traditionally underrepresented or marginalized in discussions surrounding health misinformation, achieving a truly inclusive stakeholder engagement process is critical. Our strategy includes targeted outreach efforts and the adoption of snowball sampling techniques during stakeholder interviews to widen our engagement net and ensure a diversity of perspectives is captured.

Lastly, the ethical dimensions of data collection and analysis, especially concerning sensitive health information, demands rigorous attention. Our commitment to ethical integrity encompasses the entire research process, from ensuring informed consent and anonymizing participant data to establishing a comprehensive data governance framework to guide the ethical use of data. By thoughtfully addressing these methodological challenges through targeted

strategies, we present below how to integrate findings from the various phases and methods of our study.

3.5.1 Methodological Challenges and Mitigation Strategies

Our research on creating a health fact-checker API, informed by Value Sensitive Design (VSD), culminates in an Ethical Framework that synthesizes our findings into a practical design. This framework is the backbone of the API, ensuring it embodies key human values like accuracy and privacy, adheres to ethical norms, and meets technical requirements for effective health misinformation management. Central to the framework are:

- **Value Integration:** Incorporating stakeholder values into the API to ensure it champions transparency and user trust.
- **Normative Guidelines:** Establishing operational guidelines that reflect ethical considerations and societal expectations.
- **Technical Specifications:** Defining the technical features that allow the API to robustly assess health information while being user-friendly and secure. A mock-up of the API will translate this framework into a tangible non-functional prototype, showcasing its potential and serving as a basis for iterative refinement with stakeholder feedback.

Outline of Research Methodology

This section provides a detailed breakdown of the research strategy utilized throughout the study, as outlined in Table 6. This approach is structured around three distinct phases of Value Sensitive Design (VSD), as previously outlined as the tripartite approach. For each phase, we delineate: the methods employed to gather data, the timeframe of data collection, the selection from the previous presented rationale adopted for stakeholders to be consulted within the disinformation control ecosystem, the demographic details of the participants, and the type of analysis of the data acquired.

The outline section aims to elucidate the comprehensive approach taken to achieve the research objectives already presented before. This detailed plan is divided into three VSD stages, each meticulously designed to cover specific goals, data collection techniques, data analysis procedures, and targeted outcomes to propel the research forward efficiently.

Table 6. Summary of the research used

VSD Phase	Data Collection Method	Participant Category	Date	Sample	Analysis Process
Conceptual Investigations	Systematic Reviews and Meta-Analyses	It will mention them all (Identifying, Distributing, Supporting, and Production).	Publications in the period of 2020-2022.	57 Studies included in review.	PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).
Technical Investigations	Case Studies	It will particularly mention the categories Identifying and Production.	Cases analyzed were acti.	6 case studies analyzed from chatbot solutions to 3D characters.	Descriptive analysis.
Technical Investigations	Semi-structured interview	Interview stakeholders from the Production category.	Interviews conducted between 2023 and 2024.	1 in-depth interview.	Thematic analysis.
Empirical Investigations	Co-design Workshop	Engage stakeholders from them all (Identifying, Distributing, Supporting, and Production).	Carried out from 2023 to 2024.	3 workshops held with 30 participants in total.	Thematic analysis.
Empirical Investigations	Focus Group	Interview stakeholders from the Production category.	Conducted on 5th December 2023	1 focus group of around 30 minutes with 12 participants.	Thematic analysis.
Empirical Investigations	Paper Survey	Interview stakeholders from the Production category.	Questionnaires applied on 5th December 2023.	7 questionnaires with 13 questions applied to 12 respondents.	Exploratory Data Analysis.

Only by questioning communication according to what people think and do with it can we aspire to safeguard democracy by defining the concepts of freedom of opinion and expression within the framework of communicating communication.³

(Cardoso 2023, 297)

CHAPTER 4: Technical and Empirical Results

In this chapter, we will present the results obtained from the technical and empirical methodological VSD phases of our research. In the technical research, we focused on studying cases that explored the use of AI for generative media. We aimed to understand technical aspects of these solutions that could be useful for the final proposals we intend to make in the latter part of the technique. We began by examining some AI solutions, such as chatbots for fact-checking in Canada.

Nevertheless, our technological research extended beyond this initial scope; we further explored the application of generative media technologies for summarization, exemplified by the Globo case study, to evaluate their potential for fact-checking. This part of the research centered on in-depth interviews.

While this research endeavors to provide valuable insights into fact-checking and AI development practices through a diverse array of case studies and developer interviews, it is crucial to acknowledge the potential limitations. The findings may not be entirely generalizable due to the specific contexts and characteristics of the selected case studies. Furthermore, the number of case studies and developers interviewed may be limited due to resource constraints. The primary objective of this chapter is to offer a comprehensive understanding of generative media solutions, their practices, and to evaluate the efficacy and limitations of fact-checking solutions.

³ Original: Só questionando a comunicação, em função do que as pessoas pensam e fazem dela, é que poderemos aspirar a salvaguardar a democracia, definindo os conceitos de liberdade de opinião e de expressão, no quadro de uma comunicação da comunicação (Cardoso, 2023, p. 297).

In the second part of this chapter, we sought to conduct an empirical deep dive with our stakeholders concerning the values to be integrated into the proposed solution. To comprehensively illustrate the application of these methodologies, we now present the findings from our case study analysis and interviews. This is a significant and participatory phase proposed in the value-sensitive design methodology, which, as explained in the methodological chapter, will explore some VSD methods along with focus groups and other participatory methods. The objectives of these workshops are to enable participants to critically review existing technological probes, considering their ethical implications, value alignment, and potential areas for improvement. The VSD methodology will be applied to selected probes to assess its effectiveness in identifying ethical considerations and value conflicts, with proposed refinements where necessary.

Section 4. Case studies and Interviews

In the subsequent sections of this chapter, we shall present the findings from our case study analysis and interviews, followed by an in-depth discussion and synthesis of the results. Building upon these findings, the subsequent sections address the practical challenges and methodologies involved in the publication of fact-checking results. The insights from this research endeavor is to advance our understanding of prototype suggestions, fostering more effective and efficient practices in future projects. As previously identified, the fact-checking publication phase has the potential to be explored, and we will utilize the qualitative approach to validate this hypothesis and propose an innovative solution.

Subsequently, we will explore a generative media solution using virtual characters and chatbots designed for fact-checking to offer a solution that enables these two applications of AI. In the following sections of this chapter, we will delve into the methodologies employed to conduct case studies on generative media developers and the approaches taken to gather insights from fact-checking chatbot experts. By combining evidence with theoretical frameworks, we aim to offer a comprehensive analysis that highlights the potential dangers of generative media and presents strategies to uphold truth in an increasingly complex media landscape.

Nonetheless, this research contributes to the knowledge of generative AI design practices, offering practical implications for developers, journalists, and stakeholders within the software industry

4.1 Publishing Fact-checking Results

In the context of the publication phase of fact-checking, we shall explore several cases to illustrate the practical challenges of disseminating fact-checking outcomes. Traditional methods often manually publish results using visual symbols to categorize the verification outcome as false, true, or indeterminate (Figure 26). Following our exploration of fact-checking publication examples, we shall now examine the role of generative AI and chatbots within the fact-checking process. Our research does not intend to classify existing manually created standards; therefore, we shall refrain from delving into the details of semiotics or the standards established by these initiatives.



Figure 26. Fact-check publication utilizing a visual element to denote the claim as unfounded. Source: Observador

Another example is provided by Full Fact, which has initiated campaigns to analyze COVID-19 assertions and underscore inaccuracies (Figure 27). In 2010, Full Fact became a monitoring platform with industry-specific software for claim recognition, statistical analysis, and fact-checking based on pattern recognition and structured data (Arnold 2020). Their efforts have influenced political debates

and resulted in corrections from ministers, government departments, advocacy groups, and the media. Full Fact embarked on live fact-checking with BBC Question Time, playing a pivotal role in securing corrections from the Daily Mail and the Sun (Sippitt 2019).



Figure 27. A campaign crafted to disseminate fact-checked information regarding Covid-19

Full Fact has received sponsorship to develop automated fact-checking tools like 'Live,' which instantaneously verifies statements against a database of authenticated facts (Arnold 2020).. Another tool, 'Trends,' monitors and exhibits the proliferation of misinformation. Their approach encompasses (i) direct engagement with the claimant, (ii) contacting another entity capable of influencing the claimant, such as a regulatory body, (iii) utilizing public channels to impact the claimant's stance, and (iv) adopting long-term strategies, including policy positions and campaigning for change (Arnold 2020).

In 2020, Full Fact conducted interviews with professional fact-checkers from 24 organizations servicing approximately 50 countries to gauge their perspectives on the use of technology (Arnold 2020). The findings highlight significant challenges associated with technological solutions, revealing a consensus that tools designed to automate fact verification may need to be more efficacious. Many professionals believe that the intuitive aspect of their work cannot be automated, although certain elements of their tasks can be. Concurrently, there is an

acknowledged need for strategies to aid in identifying claims, including those previously fact-checked, and locating pertinent evidence. Full Fact's workflow is structured to assist fact-checkers by enabling them to (a) monitor news websites and social media, (b) identify and categorize claims, (c) ascertain if the claim has been previously verified, and thereafter (iv) augment claims with campaigns to support the fact-checking process been already verified, and then (iv) enrich claims with campaigns to help the fact-checker (Sippitt 2019).

4.1.1 Chatbot and Fact-checking

As previously introduced, Generative Artificial Intelligence (AI) has also emerged as a tool in natural language processing, with chatbots representing a notable application. As a subset of artificial intelligence, it employs machine learning techniques to generate new content, such as text, images, or audio, akin to human-created content. One of the significant applications of generative AI entails the development of chatbots capable of engaging in conversations that mimic human interaction. To elucidate these points further, we provide an in-depth case study of the Fact Check Assistant developed by the Social Media Lab in Canada (Jo 2023). These chatbots utilize algorithms such as Recurrent Neural Networks (RNNs) and Transformer models to produce coherent responses (Jo 2023).

Generative chatbots present several advantages over rule-based counterparts. Their capability to respond to natural language facilitates more engaging and contextually relevant conversations (Denning 2023). Furthermore, they can be trained on vast datasets, allowing them to generate a wide array of creative responses. Consequently, generative chatbots are employed in customer support, language translation, and as personal assistants, offering tangible benefits to users (Adamopoulou and Moussiades 2020). Despite the promising prospects of generative chatbots, they also introduce ethical considerations requiring careful deliberation, as their training requires extensive data, often including sensitive user information, thus raising data privacy and health-related concerns (Harrer 2023).

Chatbots learn from their training data, and if said data harbor biases, these prejudices may manifest in the chatbots' responses. Such biased responses could perpetuate stereotypes, leading to unfair treatment of specific user groups (Aydiñ

and Karaarslan 2023). It is incumbent upon researchers and developers to endeavor to identify and mitigate biases in training data to ensure fairness in AI-generated interactions. Moreover, chatbots are designed to simulate human-like responses, potentially fostering emotional connections with users. Consequently, it is vital to ethically design chatbots, including disclosing their AI nature to users, to avert any harm caused by emotional dependencies (Lin et al. 2020).

Addressing the ethical challenges posed by generative chatbots necessitates a comprehensive approach. Researchers and developers must collaborate to establish explicit data collection, storage, and utilization guidelines to safeguard user privacy (Følstad et al. 2021). They should also regularly audit and monitor chatbot responses to identify and rectify biases and misinformation. Implementing safeguards against the malicious use of chatbots, such as CAPTCHAs or authentication protocols, is essential. Educating users about chatbots' limitations and their AI nature is crucial for managing user expectations effectively (Blodgett-Ford, Hennes, and Bowden 2023). Certain initiatives advocate employing chatbots in fact-checking; we will explore these models herein.

4.1.1.1 Fact Check Assistant: The Canadian case study

The "Social Media Lab's 2023 Canadian #AI Misinformation Hackathon" tackled the escalating issue of misinformation on social media platforms ("Fact Check Assistant" 2023). A hackathon is an event where participants address a specific challenge, problem, or theme, often within technology, innovation, or social concerns. The hackathon's organizers or sponsors may delineate these challenges, encompassing a broad spectrum of areas such as artificial intelligence, blockchain, healthcare, education, sustainability, and beyond.

The principal aim of a hackathon is to foster creativity, collaboration, and swift problem-solving. Participants, typically working in teams, brainstorm and bring their diverse skills to devise a functional prototype or proof of concept within the allotted timeframe (P. Johnson and Robinson 2014). Hackathons are characterized by their dynamic and intensive nature, wherein participants are fully engaged in coding, designing, and testing to realize their concepts (Yuan and Gasco-Hernandez 2021). At the event's conclusion, teams generally present their

projects to a panel of judges, with winners chosen based on criteria such as innovation, technical execution, usability, and potential impact (Poncette et al. 2020).

In response to this challenge, the Social Media Lab developed the "Fact Check Assistant" ("Fact Check Assistant" 2023), an AI-powered prototype designed to assess the viability of using artificial intelligence for fact-checking to mitigate online misinformation's effects. According to the authors, the interface is user-friendly and straightforward (Figure 28). Subsequently, we investigate another innovative project, the TITAN initiative, which seeks to enhance citizens' critical thinking skills through AI-powered coaching. The selected case studies, Fact Check Assistant and TITAN, were chosen from among several similar initiatives due to their unique focus on utilizing AI for fact-checking and enhancing critical thinking skills, respectively. These distinct approaches provide deeper insights into the potential of AI in combating misinformation, ensuring a representative and comprehensive analysis of the subject matter.

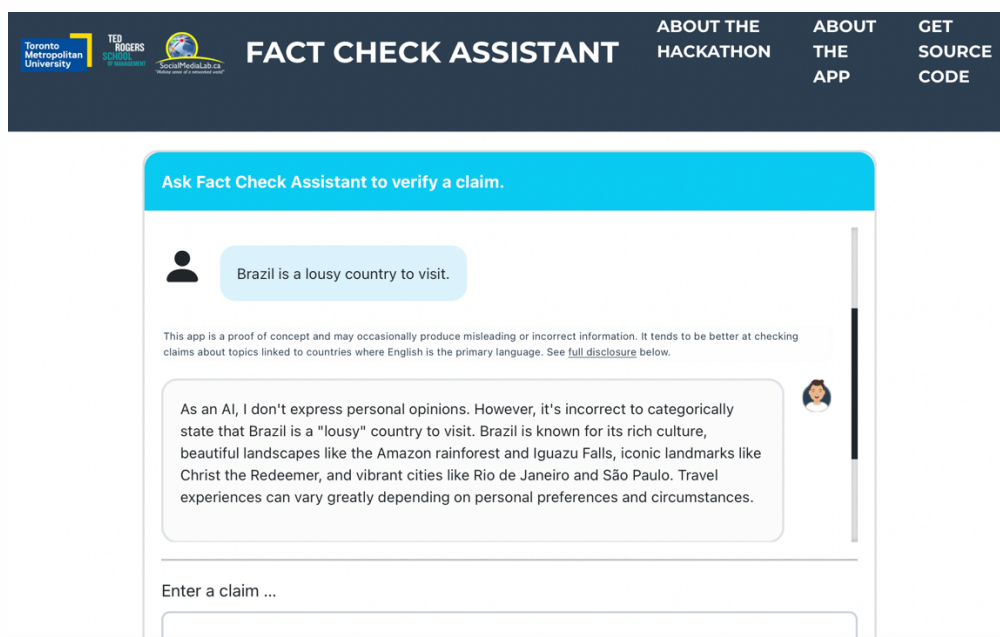


Figure 28. Fact-check assistant interface

According to the authors, the Fact Check Assistant's primary goal is to demonstrate AI technologies' potential in countering misinformation on social media platforms. The developers explain that the application considers the fact-

checking process in a simplified way, using OpenAI's advanced 'gpt4.0' model to search for verifications. However, it is imperative to note that this language model was primarily trained on English language sources, reflecting cultural values, knowledge, and biases predominantly from the English-speaking world, notably Canadian values and laws.

It is crucial to recognize the limitations of the Fact Check Assistant. As a prototype, it is provided "as is" without any warranty of accuracy, reliability, completeness, efficacy, timeliness, or the correct sequencing of information. Users are advised to exercise caution when relying solely on the app's information, as it may occasionally generate incorrect or misleading content. Furthermore, the app's knowledge base is restricted, with awareness limited to events up to 2021. The solution presents an example of employing AI to publish check results, allowing users to verify these publications automatically. From an ethical standpoint, the accuracy of the results is not guaranteed, as using GPT4 does not precisely delineate the source of the search or its inherent biases.

According to the authors, to mitigate the potential risks associated with the app's limitations, the Social Media Lab encourages users to cross-verify the information obtained from the Fact Check Assistant with additional reputable sources. This precaution is vital for ensuring the accuracy and credibility of the information presented. In conclusion, the Fact Check Assistant, according to the developers, represents an innovative step towards utilizing AI technologies to combat misinformation on social media platforms. While it serves as a proof of concept for the "Social Media Lab's 2023 Canadian #AI Misinformation Hackathon," cautious use is recommended due to its inherent limitations. Future iterations may address these limitations as technology evolves, offering more robust fact-checking capabilities. For more information and access to the Fact Check Assistant prototype, interested parties are invited to visit the project's repository on GitHub ("Social Media Lab" 2023).

4.1.1.2 Titan: AI-Powered Coaching System for Combating Disinformation

Another initiative that explores the concept of AI for the automated publication of fact-checking results is the TITAN project ("TITAN" 2023), a Horizon

Europe research endeavor. This project aims to develop an AI-powered chatbot utilizing the Socratic dialogue approach to endow citizens with critical thinking abilities. Proceeding from this, we delve into the role of virtual characters and deep learning in the context of fact-checking. Involving participants from eight countries, it addresses the limitations and possibilities inherent in such an AI-based citizen coaching ecosystem.

The principal objective of TITAN, as described by its developers, is to aid citizens in enhancing their critical thinking skills through an innovative Socratic methodology (Figure 29). Users will encounter questions to raise awareness of typical disinformation signals, such as clickbait content, manipulated images, and logical fallacies. Furthermore, the system will provide microlessons on fact-checking and AI ethics, equipping users with the essential knowledge to navigate the information environment effectively. The developers of TITAN claim that, to ensure a human-centered and ethical approach, the project incorporates citizen communities from diverse backgrounds, adhering to the European human-centric perspective on Trustworthy AI. The developers underscore the significance of impact and sustainable value alongside the creation of benchmark datasets for training AI algorithms.

1 Analyse claim

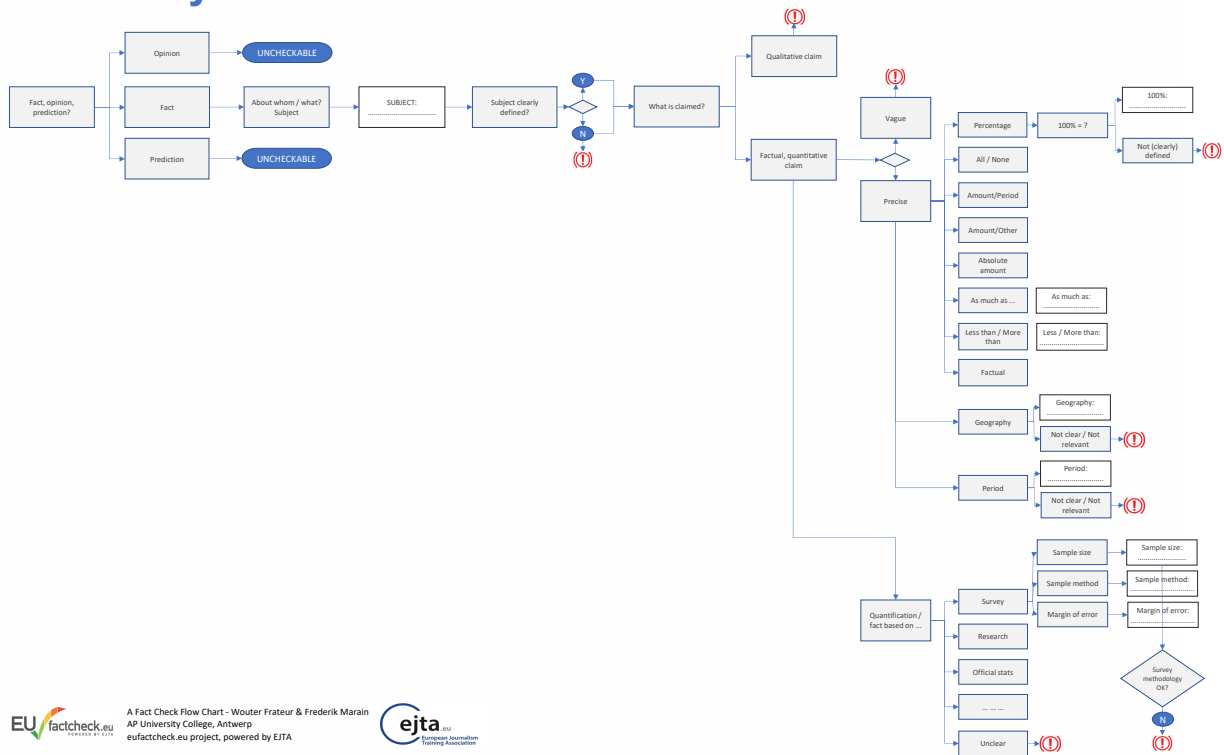


Figure 29. TITAN revision for a fact-checking flow chart

The project also develops its framework through participatory workshops. These workshops highlight the importance of AI systems in countering the spread of misinformation, particularly in light of recent scandals involving propaganda and fake news. AI-driven solutions, including fact-checking websites and social network moderation systems, have become indispensable in effectively addressing disinformation. As outlined by the TITAN project developers, the initiative seeks to pioneer the development of an AI-based chatbot employing the Socratic dialogue approach. By presenting users with pertinent questions, TITAN facilitates the identification of clickbait content, altered images, and logical fallacies prevalent in misleading news articles.

The developers assert that integrating the Socratic methodology within TITAN encourages users to scrutinize information critically, thereby fostering their logical reasoning abilities. The ultimate goal is to equip citizens with the required skills to draw informed conclusions. Beyond serving as a conversational tool, TITAN provides microlessons to users, broadening their understanding of fact-checking

and AI ethics. This strategy amplifies users' comprehension of the complexities and implications of disinformation, empowering them to navigate the evolving landscape of news and information adeptly.

In the solution proposal, the developers discuss the limitations and potentialities of TITAN and similar AI-based coaching ecosystems. The developers state that ethical concerns, such as the risk of biased responses and inadvertent collection of sensitive data, were rigorously addressed. The necessity for continual development and constant monitoring was emphasized to alleviate potential risks associated with AI systems. Key considerations of accuracy, transparency, and user privacy were spotlighted to forge a compelling and trustworthy coaching system. The developers of TITAN claim that the project exemplifies the capacity of AI to bolster critical thinking and counteract the proliferation of disinformation in our increasingly digital world.

4.2 Virtual Characters and Deep Learning

Advancing the discussion on technology and virtual characters, we explore empirical work through a case analysis of phenomena and the interpretation of their meanings. We shall delineate their technological methods, automatic content production outputs, inquiry phases, and contextualization with problematized theoretical references. Building on these technological explorations, we now present the outcomes of our co-design workshops, which aimed at evaluating technological probes and testing the proposed VSD methodology. This stage aims to scrutinize the interaction problem between technology and journalism, adopting a perspective focused on examining the complexities within a specific newsroom environment.

The case selected is from Globo, the largest television network in Brazil and one of the world's top 20 media conglomerates (Essenfelder et al. 2019). The organization has embarked on developing automated journalism systems, striving to push the boundaries of achievable media technologies and forge innovative avenues for journalism. Based in Rio de Janeiro, its team comprises software

developers, researchers, engineers, designers, 3D artists, and business professionals.

This selection emerged from a research opportunity in 2018, through my participant observation as a professional journalist within the company. This investigation, which culminated in a master's dissertation, also facilitated connections with other case studies now scrutinized in this chapter. Particular interest lies in AI and Natural Language Processing (NLP) technologies, with applications in summarization and synthetic media. The case highlights an issue related to Brazilian polls measuring voting intentions during the 2018 elections. Seizing this opportunity, the innovation team developed an NLP solution that automatically analyzed poll data and generated narratives for journalists to present on air. This phase included a case analysis and an in-depth interview with researcher Priscila Ariel, developer of Globo's computational graphics project and the AIDA and GAB project. The interview was conducted via digital channels on May 12 and June 26, 2023, through virtual meetings.

The AIDA system was conceived in response to the Brazilian political landscape in 2018. The 2018 Brazilian elections, held on October 7 (first round) and October 28 (second round), highlighted conflicting relations, and underscored a process of polarization within the country (Souza and Viscarra 2018; Almeida 2019). It was the first election wherein the presidential candidate, Jair Bolsonaro, was attacked during his campaign. Additionally, it was the inaugural instance of a candidate for the nation's highest executive office, former President Lula, being substituted due to incarceration.

As polarization and misinformation on social networks like Twitter and WhatsApp escalated (Mergulhão et al. 2018), the traditional press faced a diminishing credibility. Within this context, Willian Bonner, the anchor of Rede Globo's "Jornal Nacional," misread a technically incorrect text while announcing the results of an election poll by the Datafolha Institute. On September 14, 2018, whilst presenting a graph of voting intentions, a journalist announced:

From PSL, Jair Bolsonaro had 22% in August, oscillated within the margin of error to 24%, and now went to 26. With the margin of error, there are 24 to 28. From PDT, Ciro Gomes had 10%, then 13, now kept 13, by the margin of error, 11 to 15. Fernando Haddad of PT had 4, then grew to 9, now fluctuated to 13. With the margin of error, he has 11 to 15 (Jornal Nacional, 2018, emphasis added).

The misjudgement arose from the erroneous use of the verb "oscillates." From a technical standpoint, candidates "oscillate" whose percentage variation in voting intention falls within the survey's margin of error (Essenfelder et al., 2019). Should the results surpass the margin of error, the accurate terminology would be "ascend" or "decline," among others. Given the sensitive nature and the significant reverberation that the statement provoked, Globo team has resolved to proceed with a preconceived project: a system that generates texts for television news broadcasts from election poll data. The AIDA (Data Interpreter Algorithm) system was devised over 40 days by Globo developers to calculate the candidates' trajectories for the Presidency, identify any ties among them, and subsequently employ a text-to-text technique (Ariel 2023).

AIDA was utilized three times during the 2018 elections, exclusively on the Globonews cable channel, as deploying it on Jornal Nacional was deemed too hazardous (Essenfelder et al., 2019). Subsequently, Globo researchers have endeavoured to refine the system, aiming for a more structured automated end-to-end roundup newscast with a synopsis of the current news. To achieve their objective, AI engendered a life-like virtual character tasked with delivering a news roundup. AIDA evolved into a prototype capable of streaming in real-time across conventional media channels and the web, or creating videos for Video on Demand - VOD (Ariel, 2023).

Techniques from game engines have yielded a photorealistic character through real-time rendering, motion capture, and facial capture technologies. To render the character's behaviour more naturally, motion capture technology was applied to a real actress. The character's phenotypical features were created to honor Maju Coutinho, a Brazilian journalist who was the first black woman to anchor the Jornal Nacional (de Lima Gomes and Kneipp 2020). This event was widely reported and represented a racial milestone within Brazilian journalism in 2019. These movements, post-adjustment, were transposed to our character actress. These movements go through adjustments and are transferred to our character (Figure 30). Real-time rendering, as employed by game engines, facilitates the production of timely content, a feat not currently achievable with deepfake methods (Cowley 2018).



Figure 30. AIDA artistic process (Ariel, 2023). Source: Globo

The system operates within a pipeline comprising several stages. The initial stage involved developing an in-house data-to-text Natural Language Generation (NLG) technology to articulate the content succinctly. A text summarization engine was then formulated to generate brief descriptions of full-length news articles automatically. Following the amalgamation of both texts into a script, a deep learning-based text-to-speech system was employed to produce a natural-sounding voice with an emotional tone congruent with the news content. The character's lips were synchronized with the audio automatically and in real-time, based on the analysis of the previously generated audio. The resulting render can be tailored to different audiences by selecting news categories, virtual scenes, and relevant advertising (Ariel, 2023).

An interview with Priscila Ariel (Appendix K), the engineer responsible for the graphical computation development of the AIDA system, was conducted to glean a deeper understanding of the process. The interviewee clarifies the variations in tones and presentations, detailing the technologies presented and other work processes, the highlights of which are presented in Table 7 below.

Table 7. Highlights from the semi-structured interviews

Thematic Code	Asserts Example
Contributions of automation to productivity increases	As proof of concept, we developed a system where users can automatically create a newscast by picking options from a web page. We designed the system to be simple for a journalist or end-user. First, the user must choose what kind of tone to use with their target audience. In our system, we have two styles (or modes): one named "formal" while the other is labeled "informal" (Figure 31). Aida wears a sober outfit in the formal setting, has a serious countenance, and talks with a

neutral voice throughout the program. The scene depicts a classic newscast stand where she remains seated while a panel in the back display's illustrations relevant to the current news. On the other hand, in the informal setting, Aida is cheerful at the introduction when welcoming the user and even smiles. She stands beside a floating panel where the illustrations are displayed. Also, she wears a casual outfit and gesticulates more when talking. When designing both modes, we tried to target audiences with different preferences and tastes, making them more comfortable and engaged (Ariel, 2023).



Figure 31. Presentation tones: (a) formal and (b) informal (Ariel, 2023)

The system architecture operates in six stages. Initially, users can select from remote access options to the interface, which offers a list of current news and presentation tone style options. Subsequently, the chosen news undergoes a summarization process. In the third step, AI transforms the summarized text into audio. This is followed by lip synchronization, a process known as lip-syncing. The fifth step involves rendering the scene, where the character presents the news according to user preferences. Both steps four and five occur in real-time. Finally, in step six, a video is generated instantaneously.

The new AIDA system remains a prototype and has only been showcased at specific digital fairs, such as the Comic Con Experience (CCXP) in São Paulo, Brazil. Despite this, G1, the Globo news web portal, utilized some components of the project for journalistic endeavors during the October 2020 Brazilian municipal elections. The project facilitated the automatic announcement of elected mayors

from 5,568 cities in a report format ("G1" 2020) with all texts subject to review by a journalist prior to publication.

The system introduces the potential for integrating 3D modelling techniques into conventional domains, such as television news. The creation of avatars or virtual humans' places synthetic media as a viable option to engage audiences who no longer connect with television news as strongly (A. Brown 2018). It's noteworthy that game engines enable customization and the real-time insertion of new elements into the scene, delivering exceptional graphical quality, thanks to modern video cards. Interviewee emphasizes additional points and we highlighted some in Table 8.

Table 8. Written feedback gathered in the interview about aesthetic

Thematic Code	Asserts Example
Technology advantages	Even if it is a robot aesthetic, you still do not make a correlation with the human and then you accept the robot well, you do not have strangeness in seeing the robot, but if you put human elements there like skin, hair, a realistic human appearance, the person who is watching that character or seeing it will expect human behaviors, it will not try to look at it as a robot, it will expect human behaviors. The challenge is to bring more and more realism to the character in order to cause less strangeness when the person sees it. So it's easier nowadays to have a cartoon or an animation, than if you put a human like this here, a realistic one. This one is going to be much more rejected. 3D character is still very difficult to bring a human convincing. That is why when you want to deceive someone about a news story you use deepfake technology, because 3D is more difficult to achieve the result.

Technical applications face several critical considerations, as changing the character's attire and the scene to match the chosen mode (formal or informal) are currently limited to two options. Furthermore, synchronizing the generated speech with the character's lip movements presents limitations. The delay in execution is noticeable, and despite advancements, the voice automation's artificiality is evident. Concerning lip sync (Figure 32), the visemes—lips and facial expressions corresponding to the speech segment — utilize a neural network. This prediction, language agnostic, does not accurately cater to Portuguese. Each viseme is mapped to a geometry morph target with an expression strength (Ariel, 2023). Interviewee

notes the significance of photogrammetry for the Globo prototype and we give asserts with thematic codes (Table 9).

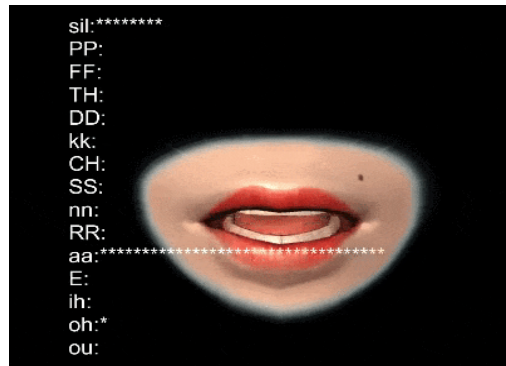


Figure 32. Analysis of the input audio stream (Ariel, 2023)

Table 9. Codes associated with our research questions and asserts from Interview

Thematic Code	Asserts Example
Contributions of automation to productivity increas	<p>The photogrammetry takes a real person and transform it into 3D. If you wanted to imitate someone by bringing them virtual, you would have to model it, bring textures. So, the 3D technology provides several ways to capture the elements, with a scanner and transform them into 3D characters. The characters are composed of a mesh: you have several dots representing each part of the object when and you connect these points you have a mesh. Then you can generate a mesh from these points, and there are the processes of optimizing the mesh so that it is not too heavy to use in real time. We are talking about a mesh character used for games, animations and these characters can gain movement, and the meta unreal is a technology that is democratizing the creation of more realistic characters. You can create with iPhone and scan a person on the iPhone to play on Meta. We have tutorials that show how to play this file. And then you can get your face on that tool and customize the rest of the body and you get your avatar down. The performance that is made here can reproduced in the characters, even with the Kinect on the mobile phone, and give movement to the 3D character.</p>
Ensuring fairness in AI applications	<p>Aida combines 3D with AI and deep learning (text to speech) to generate the voice. This exactly and this technology to generate voices is already widespread and it is very simple to put a synthetic voice with tools from Microsoft and Amazon, you already have several types of voices in several languages that you already pay a very cheap service.</p>

Additional characters, such as the gamer influencer GaB (Figure 33), a virtual presenter developed using the AIDA model and tailored for a younger audience, were created using this system.



Figure 33. GaB presents sports News. Source: Globo

The character embodies the traits of an eSports fan and player, well-informed and connected to the digital information sphere, Ariel highlights (2023). Her introduction aimed to deliver news and foster real-time interaction with the actual presenters of Globo’s sports channel:

Table 10. Written feedback in the interview with Priscila Ariel about technical challenges

Thematic Code	Asserts Example
Technology advantages	We can leave a machine in the cloud generating images of this character in real time and send only the image to the person's mobile phone by demand and if their internet is good, it will also go fast. So, when you have a simple local character directly in the app, you don't have this problem, you can put a little processing screen that downloads all the audios and all the voices that you want it to speak and then when it finishes downloading the character speaks in a fluid way. Because it already received everything before, and then it runs directly in the application. Then, as soon as it has received the audio, you no longer depend on the internet, it runs directly on your mobile.

The case study of Globo's AIDA and GaB systems underscores the innovative application of AI and deep learning technologies in journalism, highlighting both the potential and the challenges associated with these advancements. The insights gained from this technical insights provide a foundation for understanding how automated systems can enhance journalistic practices while also pointing to areas requiring further refinements, such as voice synchronization and the authenticity of virtual characters. Building on these technological explorations, we now transition to a participatory empirical approach as outlined in our methodology. This involves conducting co-design workshops with stakeholders to evaluate technological probes' effectiveness and test the proposed Value Sensitive Design (VSD) methodology. By engaging directly with users, we aim to incorporate their feedback and values into the development process, ensuring that the resulting solutions are not only technically robust but also ethically aligned and user-centric. The following sections will detail the outcomes of these workshops, illustrating the practical application of our research findings in real-world settings.

Section 5. Development of Co-design Workshops

This part of the research presents the outcomes of co-design workshops, which aimed to evaluate technological probes and test the application of the VSD (Value Sensitive Design) value scenarios methodology. It involves a diverse group of stakeholders, including researchers and experts. To ascertain the practical implications of these findings, we presented the developed technological probes for stakeholder evaluation. The workshop focus was on creating a comprehensive set of values, norms, and technological requirements to guide the design of technological solutions. The steps involved were: a) engaging participants in various participatory techniques to elicit stakeholders' values and identify commonalities and conflicts among them; b) collaboratively developing a normative framework based on the elicited values to provide ethical guidance during the design process; c) translating the identified values and norms into specific technological requirements for the probes and analysing technological probes; d) iteratively validating and refining the generated values, norms, and

technological requirements through discussions and feedback from diverse perspectives.

The workshop designated for direct stakeholders' centers on developing evaluation methods and processes to assess the fit of prototypes and solutions with the project's direct stakeholders. Activities involved include: a) collaboratively designing an evaluation framework with criteria to assess the alignment of prototypes and solutions with the identified values and technological requirements; b) introducing and exploring various fit assessment techniques, such as heuristic evaluation, usability testing, co-design methods, and experience sampling; c) giving participants the option to focus on improving a specific prototype or solution or assessing their outcomes against the established criteria; d) thoroughly discussing ethical considerations related to the evaluation process and formulating guidelines for ethical evaluation.

Similar techniques have already been used in design thinking, as demonstrated by researchers from the University of Edinburgh in partnership with the BBC (B. Jones, Luger, and Jones 2023). In a participatory workshop with journalists, examples of fake news were placed alongside examples of AI used for translation (Figure 34). In the co-design activity, journalists were invited to analyse the scenarios, discuss risks, challenges, and possibilities, and consider "provocotypes," as described by the author Dr. Bronwyn Jones.

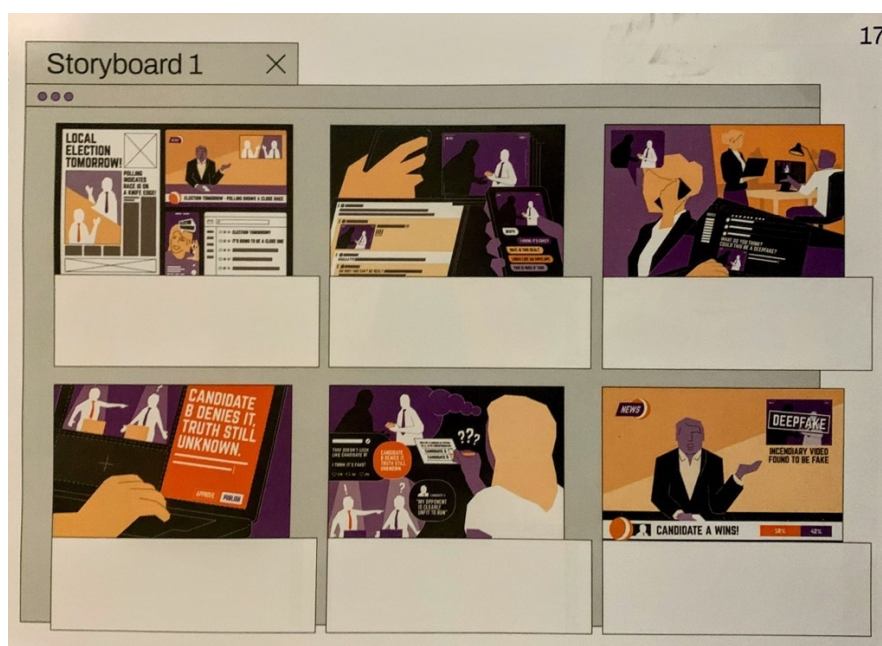


Figure 34. Storyboarding used in a participatory workshop with journalists to explore the evolution of fake news methods alongside AI translation tools.

5.1 Displaying Technological Probes

In this phase, we presented technological probes that had been developed in alignment with our objective: to use generative media to create engaging content while ensuring the accuracy of the results through fact-checking. Some of these technological probes resulted from our invitation to collaborate with the Globo media innovation team. With these probes established, we proceeded to explore value assessment scenarios through our workshops. The guidance provided was aimed at adapting the AIDA and GaB tools for efficient and swift use by stakeholders, allowing them to experience the proposed solution. Additionally, two further experiments were created by us using other APIs available on the market.

To implement this approach, we initially carried out four experiments. In the first two, we collaborated with Globo's AIDA and GaB prototypes creators to simulate how these characters would perform in a fact-checking system. For the remaining two experiments, we utilised two generative media APIs, resulting in four distinct scenarios for user evaluation. The Globo prototypes are capable of generating videos using summarisation and text-to-speech techniques. For our research, Globo's prototypes were used to simulate how this technology would summarise data and content from health fact-checking agencies, rather than summarising general news as previously assumed by the researchers. The texts simulating this summarisation were generated by us and will be detailed subsequently. Beyond the AIDA and GaB prototypes, we also incorporated the Synthesia API (Folens 2023), which utilises deepfake technology, and the D-ID API (Cleveland 2022), employing photos and deep learning. We authored the video scripts.

Synthesia offers a potent tool that employs deepfake technology to create highly realistic and personalised videos. It enables users to generate video content featuring virtual characters that can simulate the speak and behave like real individuals. By leveraging advanced artificial intelligence and machine learning algorithms, the Synthesia API allows developers to automate the production of engaging videos, applicable in marketing, e-learning, entertainment, and various other industries ("Synthesia" 2023). The D-ID API is a generative AI technology that

provides a self-service studio, API, and plugins, transforming still photos into personalised streaming videos and reducing production costs and complexities at scale. Established in 2017 and supported by top-tier venture capitalists, D-ID has been at the forefront of generative AI technology and has served clients including Warner Brothers Pictures, Publicis, Mondelez, Skilldora, and MyHeritage ("D-ID" 2023). The overall process for creating the technological probes using all these systems is illustrated in Figure 35.

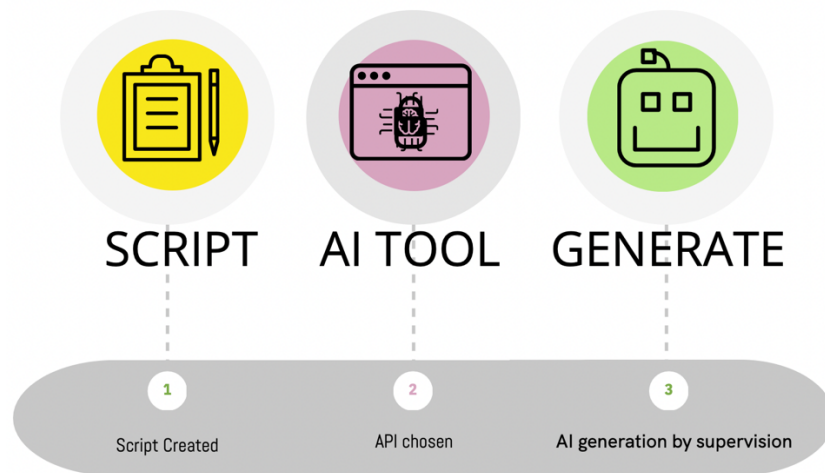


Figure 35. Basic Fact-checking Video Creation FlowChart

In the first phase, scripts were manually written for the initial three videos. A fact-checking agency was selected for each video, using a real claim from its publicly published data on official websites. Two responses—one true and one partially false—were deliberately chosen to simulate the possible outcomes of a fact check. Two references to official data on the subject were also included at the end of each video. The fourth video was a test utilising ChatGPT, wherein the system was requested to generate fact-checking content. In the second phase, APIs were chosen to automate the texts using text-to-speech technology. The first was AIDA, a 3D character from Globo; the second, GAB, also from Globo; the third, Synthesia, a deepfake solution; and the fourth, D-ID, employing text-to-speech. In the third phase, the first two videos were produced by Globo at the request of our research and collaboration project, while the remaining two videos were developed by me using public websites, free of charge on these APIs.

While it is feasible to automate the manual phases in a real proposal, our current intention was to assess the receptivity of this solution and its acceptance

within ethical values and norms; hence, only one phase was automated. Our goal was to demonstrate the potential benefits and challenges of employing synthetic media technologies to automatically generate personalised content for fact-checking dissemination, using APIs to create MetaHuman and virtual characters or to simulate different scenarios. Table 11 present the four videos created as technological probes to be used in the first two co-design workshops.

Table 11. Technological Probes presented at the first Workshops

Videos	Description	Creation process
Video 1	Fact-checking simulation using Full Fact data and AIDA character from Globo.	Script created by the research team. Video created using a pre-defined prototype already built by Globo's innovation team for AIDA.
Video 2	Fact-checking simulation using Full Fact data and GAB character from Globo.	Script created by the research team. Video created using a pre-defined prototype already built by Globo's innovation team for GaB.
Video 3	Fact-checking simulation using Poligrafo data and D-ID tool from the researcher's personal picture.	Script created by the research team. The video was created using the free tool available on the D-ID website for demonstration purposes.
Video 4	Fact-checking simulation using Full Fact data and Synthesia Deppfake character.	The screenplay was suggested when factchecking was prompt via Chat-GPT. The process can be viewed here. Video was created using the free tool available on the Synthesia.

* The videos used, as well as the demonstration of their creation process, can be viewed on Annex B.

By presenting these probes we aim to explore the aspects of contemporary society where traditional journalism is transitioning into the digital realm. This transition significantly influences stakeholders' behavior on digital platforms. The sample is not meant to be exhaustive but exploratory and only representative to some extent. We intend to delve deeper into understanding the reasoning behind these evaluations through qualitative methods. By providing co-design exercises, we aim to gain detailed insights into how digital fact-checkers meet their expectations and whether these align with the claims made by previous phases of our research.

In summary, our study employs qualitative methods to comprehensively understand digital avatars and automated audio, aiming to uncover patterns, values perceptions, and communication strategies within the digital fact-checking

landscape. At this phase of our research, we will apply a thematic analysis strategy to consider stakeholders outputs. The data presented by the workshop participants will be codified to the point where it offers us consensus to help us create a map of fundamental values that must be inserted into a solution in the final proposal to be proposed. A first round of coding will consider all the values elicited by stakeholders in their workshop exercises. In the first round we will consider all mentioned values from all participants. In a second round, comparing the workshops outcomes, we will group them into most relevant ones to guide us in a final list with codes that are derived from our data.

5.2 Workshop Unravelling Value Assessment Scenarios

The inaugural workshop was focused on evaluating technological probes and applying the proposed VSD (Value Sensitive Design) methodology. It brought together a diverse group of stakeholders, including researchers and relevant experts, and aimed to enable participants to critically evaluate the created probes to identify their ethical implications, value alignment, and potential areas for improvement. The subsequent section details the outcomes of these workshops. The workshop commenced with a 30-minute theoretical session on the VSD methodology and the exercises that would be undertaken.

A key aspect of this workshop was facilitating discussions among participants about the use of deepfake technology as a tool in journalism to combat disinformation. This provocative idea is rooted in the concepts of inoculation theory (Compton 2013), which suggests using an antidote developed from the very "virus" causing the ailment. In our context, the virus is the societal malaise known as infodemia. We aimed to engage stakeholders in considering the ethical boundaries and tangible benefits of addressing the challenges presented. Analogous to vaccine methodologies, our goal was to explore how elements of the pervasive misinformation virus might be responsibly utilized to mitigate this social ill.

This workshop took place on June 9, 2023, from 14:00 to 16:00 (WET) at the Faculty of Engineering of the University of Porto. The participants included

researchers and experts from various fields, such as journalism, telecommunications regulation, artificial intelligence, and digital media. Participants were divided into three groups, with their details provided in Table 12. 'PT' refers to participants from the testing phase.

Table 12. Participant names and details

Participant	Role	University/Company	Field of Study/Work
PT1 (Group1)	Ph.D. Candidate	University Fernando Pessoa, Portugal	Fact-checking
PT2 (Group1)	Ph.D. Candidate	Faculty of Engineering of University of Porto	Telecom regulation
PT3 (Group1)	Journalist	Freelancer, Portugal	Reporter
PT4 (Group2)	Ph.D. Candidate	University Fernando Pessoa, Portugal	PhD student
PT5 (Group2)	Ph.D. Candidate	Faculty of Engineering of University of Porto	AI and Arts
PT6 (Group3)	Ph.D. Candidate	Faculty of Engineering of University of Porto	Digital Media
PT7 (Group3)	PhD Professor	University Fernando Pessoa, Portugal	AI and Journalism

This table includes the participants' codes, roles, affiliations, and fields of study/work, highlighting the workshop's diversity of knowledge and perspectives. The participants were presented with a 30-minute theory session to initiate the session, which explained the methodology and exercises they would engage in. The workshop consisted of two main exercises (Appendix B):

a. Exercise 1: Exploring Deepfake and Deep Learning Use Cases

In this activity, within 10 minutes, the participants discussed and analyzed deepfake and deep learning scenarios related to fake news promotion. Two specific cases were presented - one featuring a deepfake video of former President Obama aimed at raising awareness about deepfake false information, and the other demonstrating the manipulation of a video of Nancy Pelosi to make her appear intoxicated. Additionally, the participants viewed videos showcasing these technologies' use in journalism, including AI anchors from Kazakh's Atameken Business and Brazilian network TV Globo's 3D virtual news anchors named "AIDA" and "Gab."

b) Exercise 2: Aligning Technology with Values, Norms, and Requirements

In the second exercise, the participants had 30 minutes to explore the alignment of synthetic media technologies, such as deepfakes, with the values, norms, and requirements identified in the previous discussions. The focus was on discussing technological probes that offered potential solutions to the ethical implications of

using deepfakes in journalism. During the group work phase, participants actively engaged with the workshop materials and used Mural's digital platform for interactive collaboration. This online platform allowed each participant to access a virtual whiteboard to discuss the tasks at hand and contribute their insights.

The workshop followed a structured agenda, and participants were engaged in a group discussion, identifying important values, norms, and requirements for evaluating the use of deepfakes in journalism. They shared their experiences, opinions, and perspectives on the topic by the end with the identified values, norms, and requirements. Besides the Mural instructions, participants could access Value Cards, a deck tool used by VSD methodologies to help people relate to existent values (Figure 36). The workshop concluded with a group pitching and a comprehensive discussion of the ethical implications of using deepfakes in journalism. Participants summarized their findings and shared their key takeaways from the workshop. The purpose of this test model was to fine-tune the workshop, identify areas of improvement, and prepare for its future applications.



Figure 36. Value Cards used in the workshop as a tool from VSD

5.2.1 The Workshop Outcomes

In this section, we will try to transcribe the comments and annotations made digitally in Mural. We will also transcribe the remarks (Appendix I) of the final pitching of each group with the participants' authorization. Some tables will also be generated to help record the comparisons and other insights. The texts have been proofread to correct grammatical errors and improve readability. The findings from these workshops provide a comprehensive understanding of the participants' perspectives and the effectiveness of the proposed technological solutions. However, it's essential to note that the text provided is a transcript of spoken discussions, and some sentences or phrases may still be fragmentary or incomplete.

In Exercise 1, Group 1 participants discussed the impact of modifying media and the potential damage caused by altering the form of communication. They emphasized the importance of authentic information and the subtlety of the message changes. They also identified technical requirements, such as social networks' potential for amplifying damage and distinguishing between live and recorded communications. Exercise 2 discussed using virtual characters for fact-checking and identified central values and norms related to public interest versus private interest and fast response versus precision and innovation. They proposed a 3D real-time fact-checking prototype based on machine learning.

In Exercise 1, Group 2 participants reflected on the ethical implications of using AI to generate fake videos, particularly in election times. They raised questions about morality, ethics, and the control of imagination. In Exercise 2, they identified central values and norms related to freedom, tolerance, innovation, and rationalism. However, they did not explicitly analyze tensions related to these values. In Exercise 1, Group 3 participants provided notes on various cases of manipulated videos and discussed the potential displacement of journalists by virtual characters. In Exercise 2, they identified central values and norms related to respect, independence, reliability, and responsibility in fact-checking. They suggested that technology should be used in the process, but a real person should deliver the information to establish a more human and close relationship with the audience.

Overall, Group 1 provided more detailed analysis and specific technical suggestions, Group 2 emphasized ethical concerns, and Group 3 highlighted the importance of human interaction and reliability in fact-checking. Each group brought unique perspectives to fact-checking and using technology in media manipulation. Nevertheless, they were convinced to use a solution with deep learning, but one that utilized a 3D character and not a deepfake of a real person. They were concerned about trust in the solution and worried about usability and whether the user would be interested and how. Therefore, they proposed that this solution be something like a translation plugin, which, with a click, would identify possible fake news, and in a popup, the check would appear in the form of a 3D character.

It was possible to realize that the groups did not have much time to finalize the last tasks, so doing only one would be better for a second application. They could also emphasize more on the discussion about values and norms (Table 13) so we decided to remove the first exercise for a second application.

Table 13. Values and norms presented by each group

	Group 1	Group 2	Group 3
Value	<ul style="list-style-type: none"> • Public health • Individual liberty • Justice • Equality • Access to information • Freedom of expression • Democracy • Innovation 	<ul style="list-style-type: none"> • Freedom • Tolerance • Innovation • Rationalism 	<ul style="list-style-type: none"> • Respect • Independence • Reliability
Norms	<ul style="list-style-type: none"> • Public interest vs. private interest: Innovation is constantly pushing regulatory boundaries. • Fast response vs. precision and innovative production. • State using technology as a deception strategy. 	<ul style="list-style-type: none"> • Making decisions based on facts, not only arguments. • Encouraging critical pedagogy and various forms of dialectics. • Fostering media literacy. • Having emotional intelligence to discern reality from the virtual. • Having an open mind for innovation. 	<ul style="list-style-type: none"> • Treating everyone equally • Fact-checking based only on verified information • Technology being stable, reliable, and secure • Having a responsible approach

5.3 VSD Methods in a Multi-Cultural Workshop

In our quest to engage a more heterogeneous array of stakeholders, we iterated on the methodology previously applied, although we commenced with a comparable structure for proposing a new workshop. The decision was made to refine our approach using similar exercises. This revised co-design workshop model was accepted for presentation at the C+J – DataJ 2023 Conference, a joint event that merged the Computation + Journalism and European Data & Computational Journalism Conferences. The conference was hosted at ETH Zurich, Switzerland, from June 22 to 24, 2023.

With a focus on the influence of AI in the media sector, the conference featured a variety of talks, workshops, and panels that addressed topics including the future of AI in news, digital ethics, natural language processing (NLP), algorithmic accountability, data journalism, and newsroom automation. Researchers from around the globe attended, offering an opportunity to explore some of the challenges highlighted in the thesis. This unique conference promoted a multidisciplinary dialogue across information, data, social, and computer sciences, aiming to advance research and practice in Data and Computational Journalism. The programme included a blend of academic presentations and keynotes from industry leaders.

Throughout the event program (Appendix C), the importance of ethical frameworks, such as VSD, in evaluating emerging technologies and combating disinformation was also emphasized. The potential benefits and challenges of using synthetic media technologies for automated fact-checking dissemination were thoroughly explored in many other initiatives. By leveraging the insights gained from the previous test model, workshop two was well-prepared for the upcoming Computation + Journalism European Data & Computational Journalism Conference 2023. The knowledge and experiences acquired in this test model have undoubtedly contributed to the success of this workshop's objectives and discussions. An agenda was previously sent to participants as a camera-ready version (Appendix D).

The workshop took place on June 22, 2023, at Tamedia, a communication group in Switzerland that hosted a complete program of workshops in Zurich. The

workshop was initially planned for 12 participants but ended up with 11 participants (Table 14), divided into three groups (two groups with 4 participants and one group with 3 participants) based on their areas of knowledge. PW refers to Participants from the second Workshop phase.

Table 14. Participant names and codes with details

Group 1	Country	University/Company	Field of Study
PW1	Spain	University of Santiago de Compostela	Journalism Studies: fact-checking
PW2	Netherlands	University of Amsterdam	Media/Journalism/Technology
PW3	Netherlands	University of Groningen	Mis-and disinformation studies
PW4	Denmark	University of Southern Denmark	Computer Science (Cryptography)
Group 2	Country	University/Company	Field of Study
PW5	Serbia	TX Group	Artificial Intelligence
PW6	España	University of Malaga	Journalism studies
PW7	Hong Kong	Charles University	Communication and Philosophy
Group 3	Country	University/Company	Field of Study
PW8	Netherlands	University of Amsterdam	Journalism studies: New media
PW9	Netherlands	University of Amsterdam	Information Law / Communication
PW10	UK	The Times	Journalism
PW11	Switzerland	University of Zürich	Comparative Language Science

For this version, participants were directed to develop a single revised exercise (Appendix E) Once again, the participants watched the four technological probes videos with Fact-checking presented by AIDA, GAB, Synthesia deepfake solution, and D-ID version. They also had the Values Cards available for consultation and envisioning cards, a method from VSD (Figure 37), as one of the tasks.

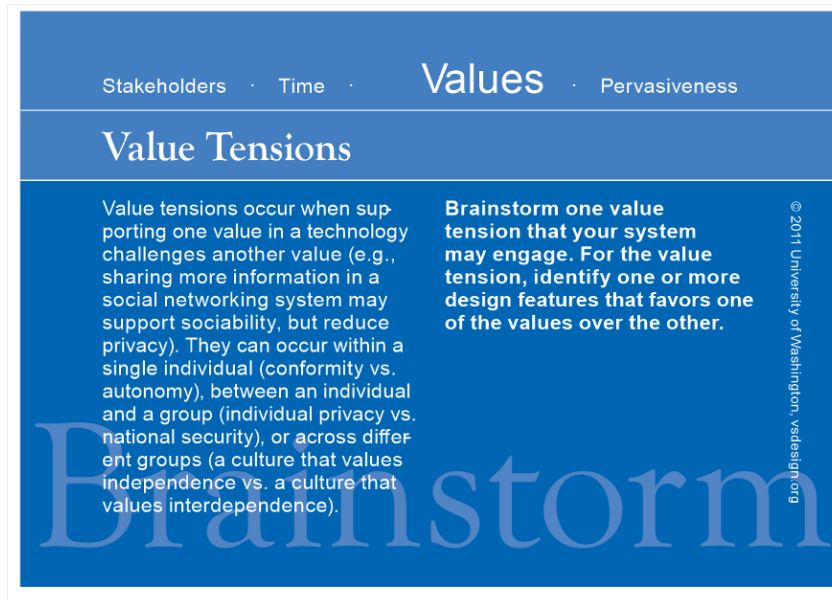


Figure 37. Envisioning Card used in the workshop as a tool from VSD

The participants were challenged on the following statement:

Tamedia plans to create a digital news verification system using deep learning technology to provide fact-checking through a news feed. They have tasked your team with evaluating the feasibility of this solution's virtual interface using VSD methodologies. Goal: Discuss how fact-checking technology can be aligned with the values, norms, and requirements identified in the previous discussions. Gather in small groups and discuss the impact of deepfake and virtual characters to propose automatic fact-checking dissemination. Each group had 30 minutes.

The workshop was initiated with a 10-minute introduction, where the facilitator presented the objectives and provided an overview of the value-sensitive design methodology. Participants were then divided into groups and instructed to take notes (Figure 38). In the following 20 minutes, the facilitator explained various technologies related to synthetic media, deep learning, deepfake, and virtual characters. Advantages and differences between these technologies, especially deepfake, were discussed. The presentation (Appendix J) included the advantages and disadvantages of technologies used for virtual 3D characters and an explanation of the photogrammetry process. This approach aimed to ensure clarity on the technical feasibility of the tools, particularly focusing on practical aspects such as adaptability and potential limitations like internet dependency, and the technical expertise required. Additionally, cases involving news agencies that utilized deep learning technologies with newscast avatars were presented. The

workshop also covered a deepfake case involving President Kim Jong Un, which was shared on YouTube to highlight the dangers of deepfakes.



Figure 38. Presentation of the initial concepts of the workshop

Participants followed a structured data collection process using sticky notes and manual writing. Alongside oral instructions, visual remarks on the left-hand side of the work areas guided them through the exercises. Each group was given the flexibility to either write their ideas on sticky notes or directly on paper, depending on their preference. The process was organized into several exercises. First, participants discussed and identified key insights, which were either noted on sticky notes or written manually. They then organized these notes into predefined categories such as influencing benefits, harms, and proposed solutions, with the sticky notes or written ideas placed in designated sections on large sheets of paper.

Throughout the process, we were a facilitator to present to clarify any doubts and provide support. However, participants were not subjected to strict time constraints, allowing for a more relaxed and collaborative environment. This flexible and supportive approach ensured that all groups could effectively capture their discussions while maintaining a consistent structure across the exercises, facilitating later comparison and analysis.

During the workshop, we ensured that each group had the opportunity to explore the core elements of values, norms, and requirements through structured tasks that encouraged discussions. The three groups engaged in five exercises, each

focusing on different aspects of the fact-checking technology and its implications. Group 1 explored the benefits of fast, visual content with a focus on societal norms, while Group 2 concentrated on diversity and the cost-effectiveness of the solution, and Group 3 expressed resistance to deepfake solutions due to concerns about authenticity and trust.

The three groups engaged in five exercises, each focusing on different aspects of the fact-checking technology and its implications. Each group approached the exercises with varying focus: Group 2 dedicated more time to balancing tensions between ethical considerations and practical benefits, Group 3 focused on proposing alternative solutions due to their reservations about deepfakes, and Group 1 did not fully develop the aspect of norms. Despite these differences, the structured tasks allowed all groups to contribute meaningfully to their discussions.

The next segment, lasting 40 minutes, focused on aligning synthetic media technologies for fact-checking. Participants received a brief explanation of the challenges and problems of fact-checking dissemination. The fact-checking process was described, highlighting the typical steps involved in verifying factual information. Studies from Full Fact Org were referenced, presenting the automatic phases in four subtasks related to fact-checking. Participants engaged in group discussions on how these technologies could be aligned with the values, norms, and requirements identified in previous talks. The facilitator demonstrated the focus on technological probes, providing insight into how these could be applied within the technology. Each group prepared their value-sensitive design ideas for this solution.

The workshop concluded with a 20-minute evaluation and conclusion session. Each group had a 5-minute pitching opportunity to present their ideas concisely and persuasively. Criteria were set for the presentations, with each group allotted up to one minute for their main points, followed by a brief Q&A session for clarification or elaboration. The session ended with a summary of the feedback, a thank-you message, and an invitation for participants to stay in touch.

5.3.1 Workshop Findings

In the five exercises developed in the workshop, Group 1, comprising participants from Brazil and Denmark, demonstrated a keen interest in understanding stakeholders and envisioning the future. Group 2, consisting of participants from diverse continents, highlighted the strength of diversity in their discussions. Group 3 exhibited resistance to deepfake solutions, associating them with disinformation.

The outcomes of task 1 centered on unpacking the benefits and harms of using avatars to disseminate fact-checking information (Table 15). In the analysis of the workshop findings, we distinguish between harms/benefits and values/norms. Harms/benefits refer to the immediate effects and consequences of the use of the technology, highlighting both the positive and negative impacts it may have on stakeholders. For example, a benefit might include increased accessibility or speed, while a harm could refer to concerns around trust or authenticity. On the other hand, values/norms represent the fundamental ethical principles and societal standards that should guide the design and application of the technology. These include broader concepts such as privacy, inclusiveness, transparency, and trustworthiness, which participants deemed essential in ensuring the responsible use of AI in fact-checking.

Table 15. Benefits and Harms for the solutions from each group

	Benefits:	Harms
Group 1	<ul style="list-style-type: none"> • Fast to consume • Visual • Short • Sharable • Fast to produce • Cheaper than a newsroom 	<ul style="list-style-type: none"> • Hard to rely on it because it's fabricated. • Easy to replicate. • No human touch
Group 2	<ul style="list-style-type: none"> • Less expensive, cost, audience, speed. • Diversity of characters. 	<ul style="list-style-type: none"> • Create more fake content
Group 3	<ul style="list-style-type: none"> • Cheap • Fast • Teach • Accessibility • Personalization 	<ul style="list-style-type: none"> • Compromising authenticity and trust. • Adding cost resources by design (in comparison to text). • Ethical concern. • Inspire more people to use deepfake. • No transparency for the audience.

We will present these results through the lenses of the groups, each contributing valuable perspectives and discussions. Participants from Group 1 discussed the advantages, such as the content’s fast consumption, visual appeal, and shareability. On the downside, there were concerns about relying on fabricated information, easy replication, and a lack of human touch in the content. The second group explored how this approach could compromise authenticity and trust but could also be cost-effective, fast, and accessible. The third group, however, faced considerable resistance to the proposed solutions from the start. Task 1 for them expressed concerns about compromising authenticity, ethical issues, and challenges in inspiring people to use deepfakes. On the other hand, they recognized the benefits of cost-effectiveness, speed, and accessibility.

We can see that all three groups highlighted the cost of producing this form of communication presented as a benefit. Two groups also highlighted personalization as an outstanding advantage and the diversity of characters as a point to note. On the other hand, the strangeness of using something “fake” to combat fake news was also an interesting point of discussion. PW5 highlights that “on potential harms, there is a possibility this will be more fake, as team one said, that maybe it will sound even less reliable to the facts. It will discredit all the fact-checking and all the work the journalists do and on the other side”. Group 3 discussion agreed:

Table 16. Asserts from group 3 discussions about harms and benefits

Theme	Participant	Asserts Example
Balancing Authenticity and Accessibility	PW9	So, I think many things you also find. In our first discussion, we had an internal struggle with bringing deepfakes together with fact-checking. So, we had to devise a way to see the benefits, but then we also identified that it’s cheap. It’s fast. You haven’t an opportunity to reach more people and more people who usually use other modes of getting the news or getting content. So, what we identified as a benefit is accessibility also for people who maybe cannot read, for example. Then they have the opportunity to see videos about that, which would be a good thing that we have here, and also cost-wise when you think of as an NGO, for example, produces video they don’t have the resources maybe to pay or higher some human to present the things. It’s cheaper, so they must at least make a video. However, they may want humans, and in the case of harm, well, with the problem with authenticity, of course. So, is it authentic when the fake person tells us something about fact-checking that’s a generic conflict here? I guess, and we have, like.

5.3.1.1. Central Values and Norms

Moving on to Task 2, the group participants were asked to write down central values and norms related to their project (Table 17). The first one listed their values as beauty, engagement, and meaningfulness. Additionally, societal norms conflicting with mental health and inclusion were analyzed. The second group identified central values and norms, such as authenticity, trustworthiness, efficiency, inclusion, and informativeness. They considered accessibility, inclusion, and repetition crucial and discussed the need for different plugins. The third group identified authenticity, trustworthiness, efficiency, inclusion, and informativeness as essential values, and they discussed the importance of transparency and various technical implementations.

Table 17. Values and norms presented by each group

	Group 1	Group 2	Group 3
Value	<ul style="list-style-type: none"> • Beauty • Engagement • Meaningfulness 	<ul style="list-style-type: none"> • Privacy • Ease of use • Transparency • Inclusiveness 	<ul style="list-style-type: none"> • Authenticity • Trustworthiness • Efficiency • Inclusion • Informativeness: Explanation, Proof, Sources
Norms	—	<ul style="list-style-type: none"> • Watermarking 	<ul style="list-style-type: none"> • Something related to accessibility, inclusion, and repetition. • Different plugins

Group 1 contemplated the role of values in the static construction of this virtual character. Since they had chosen beauty as a core value, this criterion rendered it impossible for them to employ characters with humanoid features. PW3 expressed their rationale: “We selected the following values: beauty, engagement, and meaningfulness. Therefore, this deepfake appears highly inauthentic to us. Consequently, we don’t believe it’s appealing. People on Instagram or TikTok, for instance, seek aesthetically pleasing content. That’s why we opted for a cat; perhaps a cat would be suitable, as cats are beautiful, engaging, and appealing.”

In Task 3, Group 1 presented the tensions associated with their chosen value of beauty and its impact on engagement. PW4 also explored how adhering to

societal norms could conflict with mental health and inclusion: “It doesn’t come across as insincere, but it does have the potential to inadvertently reinforce societal beauty standards, prompting individuals to conform to those standards in the pursuit of looking good. This inherently clashes with mental health concerns and similar issues.”

5.3.1.2. Thematic grouping of workshop data

At this point in our research, we carried out a secondary thematization to analyze the data collected. While themes represent the final outcome of thematic analysis, they also establish a framework that delineates the focus and scope of the analysis. Thus, following an initial phase where themes suggested by our research questions were used, the second phase utilized the outcomes of the first two workshops as our input guide. Pre-determined themes can direct the coding process and ensure the analysis remains aligned with the research objectives. Here, we followed the stages of thematic analysis in this manner: (a) data from the workshops were transcribed and we became familiarized with them, (b) values were transcribed and grouped into codes, (c) initial sub-themes were mapped (see Figure 39), (d) sub-themes were reviewed and matched with codes, (e) names were refined, and (f) reports were generated and will be presented subsequently.

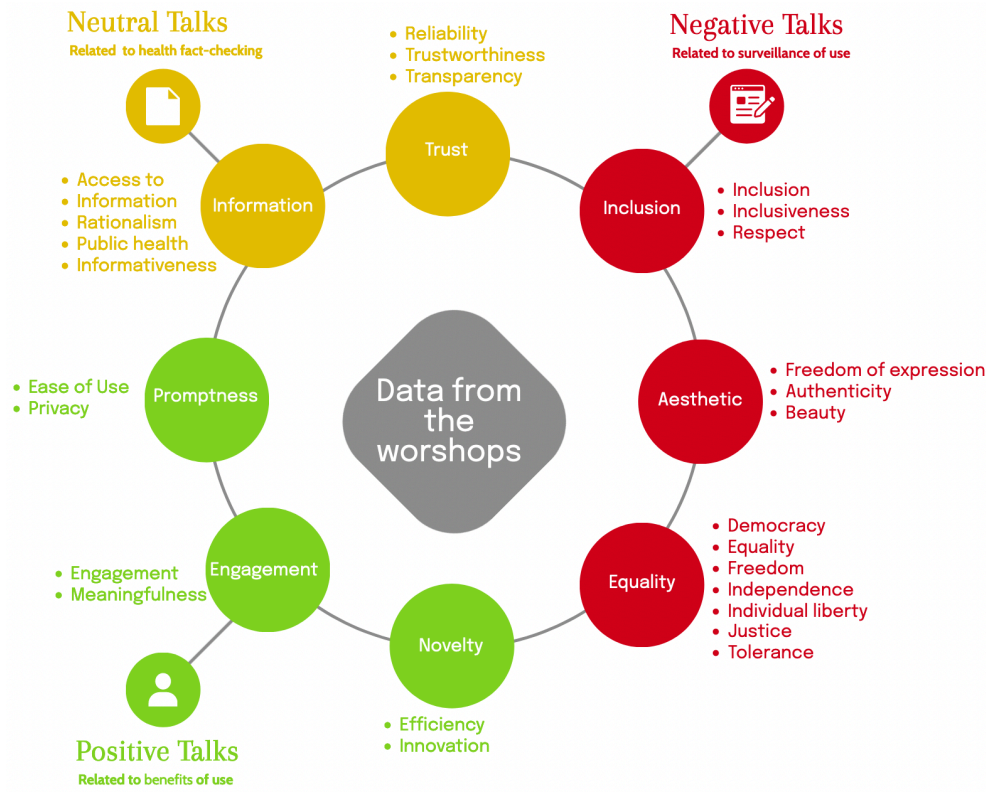


Figure 39. Developed thematic map, showing three main approaches and eight codes

Upon reviewing the outcomes of the exercises outlined, we extended the coding of exhibitions and presentations in an initial round. In this phase, we identified words considered fundamental values to include in a solution akin to our proposal. Consequently, the data directly informed the codes generated in this preliminary round. This coding is crucial for analyzing the participants' speeches during their group presentations and other focus group activities we conducted.

In the subsequent coding phase, leveraging this dataset, we will organize it into categories within Table 18 that align with their respective descriptions. This process involves constructing a grid comprising categories and subcategories, integrating the key values essential for the forthcoming stages of our thematic analysis applied in this research.

Table 18. Participant names and codes with details

Research Question Scopes	Values from Workshops	Subthematic Codes	Sub-themes
Health Misinformation	<ul style="list-style-type: none"> • Access to • Information • Rationalism • Public health • Informativeness 	Information	<ul style="list-style-type: none"> • Accuracy of health-related information • Clarity and comprehensibility
	<ul style="list-style-type: none"> • Inclusion • Inclusiveness • Respect 	Inclusion	<ul style="list-style-type: none"> • Digital Accessibility • Representation of diverse groups
Ai-Ethics	<ul style="list-style-type: none"> • Democracy • Equality • Freedom • Independence • Individual liberty • Justice • Tolerance 	Equality	<ul style="list-style-type: none"> • Disclosure of processes and policies • Ethical data use and decision-making • Representation of diverse groups
	<ul style="list-style-type: none"> • Freedom of expression • Authenticity • Beauty 	Aesthetic	<ul style="list-style-type: none"> • Cultural sensitivity • User experience design
	<ul style="list-style-type: none"> • Engagement • Meaningfulness 	Engagement	<ul style="list-style-type: none"> • Participation and involvement • Interactivity mechanisms
Fact-checking	<ul style="list-style-type: none"> • Reliability • Trustworthiness • Transparency 	Trust	<ul style="list-style-type: none"> • Availability and accessibility • Reliability and consistency
Artificial Intelligence	<ul style="list-style-type: none"> • Efficiency • Innovation 	Novelty	<ul style="list-style-type: none"> • Creativity and originality • Unique features
	<ul style="list-style-type: none"> • Ease of Use • Privacy 	Promptness	<ul style="list-style-type: none"> • Speed of response • Timely delivery of services

5.3.1.3. The Impact of AI Fact-checking Virtual Anchors

Task 4 involved identifying the best technological solution for their project and suggesting technical requirements. They considered using a famous meme

character, visual content, meme-driven strategies, and avoiding deepfakes and anchors. Through their pitching presentation, task 5, we will highlight other insights from this.

a. Group 1: Balancing Beauty and Societal Norms

This group grappled with the tension between conforming to societal beauty standards and the potential harm this might pose to mental health and inclusivity (Figure 40). Table 19 compare their thoughts.

Table 19. Values and norms presented by each group

Sub-themes	Participant	Asserts Example
User experience design	PW3	Deepfakes appear inauthentic to us, lacking appeal. People on platforms like Instagram or TikTok seek aesthetics, and 3D characters fall short. Maybe a cat would work, as cats embody beauty, engagement, and appeal.
Cultural sensitivity	PW2	If we use a top model as an anchor, we risk reinforcing the societal beauty standard, driving people towards surgeries and Botox. That's a significant ethical dilemma.

Group 1's suggestion involves visual content and MEME-driven strategies to maintain engagement while ensuring authenticity. This group highlights the importance of aligning AI ethics with chosen values.

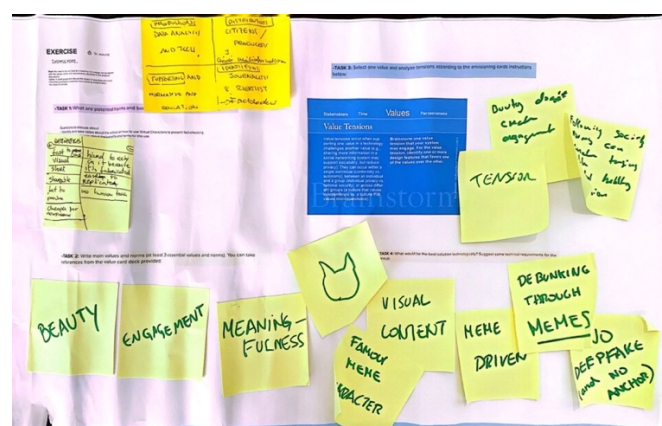


Figure 40. Group #1 experience

b. Group 2: Diversity and Ethical Considerations

This group, comprised of participants from three continents, embraced diversity in their discussions (Figure 41). They examined the potential benefits and harms of AI-driven content, particularly deepfakes (Table 20).

Table 20. Insights about diversity and ethics

Sub-themes	Participant	Asserts Example
Ethical data use and decision-making	PW4	Deepfakes may increase the prevalence of fake content, undermining fact-checking efforts. However, they are cost-effective, appealing to diverse audiences, and capable of including thought leaders. We must use them cautiously and avoid malicious use.
Reliability and consistency	PW5	It can also include people from books or some characters that we ethically and morally strive to be more, and of course, the speed of use is much higher. Fundamental values are privacy, inclusiveness, ease of use, and transparency. We must ensure users find it easy to distinguish AI-generated and human-generated content. Transparency is crucial to building trust.
Clarity and comprehensibility	PW6	So, we wanted to have the same thing as presented by the technological probes but just with installed values of higher inclusiveness and higher transparency and having it known is it AI or is it person so it can be like this, but it needs to be clear.

Group 2 aims for inclusiveness and transparency in their proposed solution, emphasizing the importance of ethical considerations and precise identification of AI-generated content.

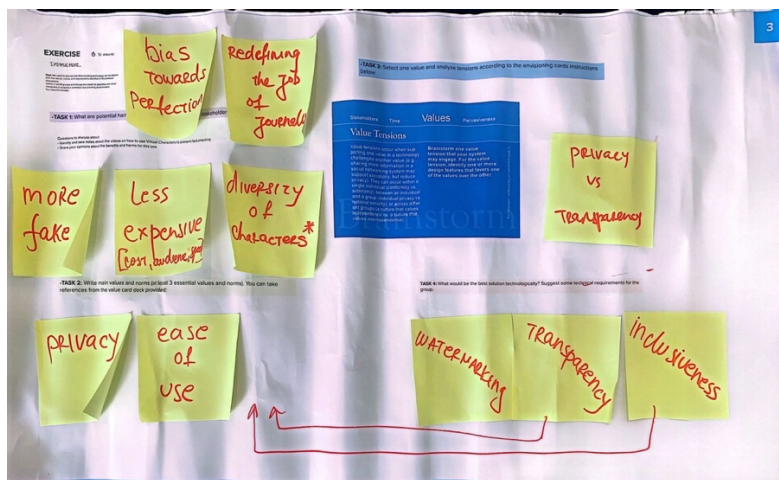


Figure 41. Group #2 experience

c. Group 3: Ethical Values and Resistance to Deepfakes

This group expressed skepticism about using deepfakes due to their association with disinformation (Figure 42). They identified various ethical values and discussed potential tensions. Table 21 delve into their perspective on the use of this technology.

Table 21. Asserts concerning data and promptness

Sub-themes	Participant	Asserts Example
Ethical data use and decision-making	PW9	We had an internal struggle regarding deepfakes and fact-checking. While they offer accessibility and cost benefits, authenticity remains a concern. The line between deepfakes and 3D models also needs clarification.
Timely delivery of services	PW11	We prefer alternatives like audio, animations, or text over AI-generated content. These formats can be cost-efficient and provide better context. Emphasizing the source and context of information is crucial.

Group 3 leans towards caution and explores alternative content formats to address their ethical concerns. They highlight the importance of authenticity, transparency, and context in conveying information.

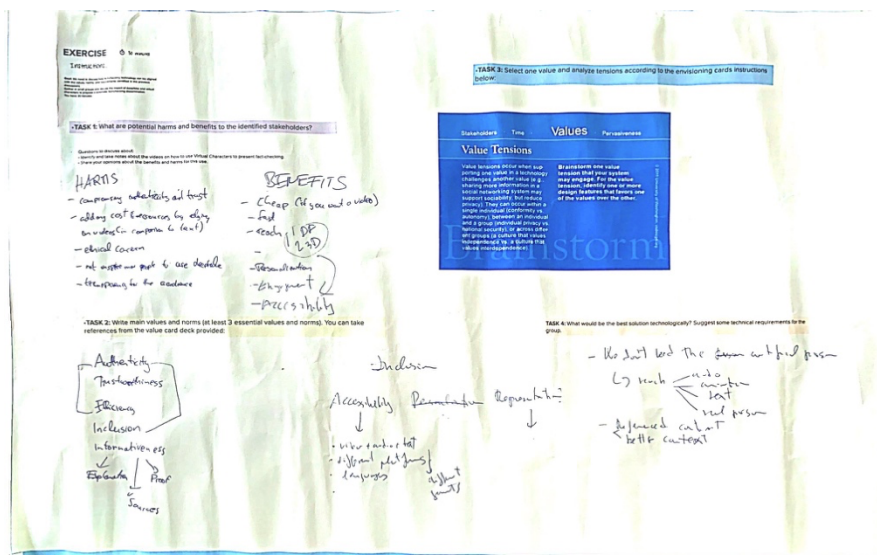


Figure 42. Group #3 experience

5.4 Value-oriented Workshop in Elderly Health Information

This new workshop was designed to run in parallel with an existing event in the Educage project called "multiplayer," which presents conclusions to the outreach community and, for this one, would bring the results of a portal realized in Moodle technology format and which presents information content on healthy aging. As one of the phases of this event, we organized a co-design workshop inspired by solutions for a fact-checking chatbot within this web portal. Based on the previous field research phase, we recognized the need to refine the participatory approach and bring in different methods, as already underpinned in the methodological section of this document.

Therefore, our methods encompassed a series of structured activities aimed at enhancing the design's impact and relevance. It began with a Questionnaire Evaluation, where participants provided feedback on chatbot prototypes through a 13-question survey responded on a Likert scale. The Focus Group Discussion, based on Krueger's (2014) method, involved structured debates to develop an ethical framework for the design process, utilizing a script to probe deeper into participants' preferences and value evaluations. Lastly, in the Brainstorming session, participants used the Empathy Map (Siricharoen, 2021) to create detailed personas, enriching our understanding of user motivations and guiding user-centered design decisions.

This workshop iteration aims to evaluate the technological solution with a focus that moves away from purely ethical considerations and towards a value-orientated solution proposal. The event took place on 5 December at the professor's house at Senior Citizens University ("Macedo De Cavalheiros" 2024), from 9am to 12pm, in the town of Macedo de Cavalheiros, Bragança. The proposed agenda (Appendix F) for the event was aligned with the overarching goals of the EducAGE project and publicized. The research involved 12 participants from the town of Macedo de Cavaleiros, ten women and two men, with an average age of 59 (Table 22). The workshop was attended voluntarily by students from the Senior Citizens University, which provides courses and activities in the city on a regular basis. The elderly caregivers' educational levels also vary, with some having basic education, others

with secondary education, and a few with higher education. PQ refers to Participants from the Questionnaire phase.

Table 22. Demographic Profile and identifications of Study Participants

id	SEX	Age	Education
PQ1	male	64	Secondary Education
PQ2	female	80	Basic Education
PQ3	female	55	Secondary Education
PQ4	female	67	Secondary Education
PQ5	female	23	Higher Education
PQ6	female	49	Higher Education
PQ7	female	55	Secondary Education
PQ8	male	69	Basic Education
PQ9	female	67	Secondary Education
PQ10	female	77	Basic Education
PQ11	female	43	Higher Education
PQ12	female	62	Secondary Education

5.4.1 Disclosing Iterated Technological Probes

At this stage of our research, the technological probes were iterated further in line with insights from inoculation theory, as we wanted to use technologies that have been closely adopted for sharing disinformation to understand with stakeholders’ ethical ways of using them to contain disinformation. For this part, then, with a focus on developing an automatic fact-checking plugin for a web portal, we tested a few advanced technologies, with special emphasis on the use of large language models (LLM) and their differentiated applications and software. The insights gained from these iterations informed the subsequent phases of our research, particularly in refining the technological solutions. The overall process of creating the technological probes using all these systems is illustrated in Figure 43.

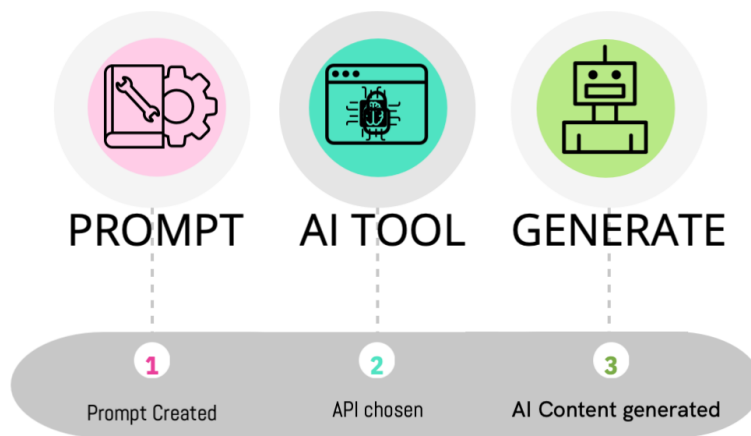


Figure 43. Iterated flowchart for creating fact-checking videos

In the initial step, prompts (Appendix G) were crafted to specify the instructions for generating the solutions to be automated by APIs. One solution used for comparison was a video produced without AI, following the traditional formats of SIC, a journalistic company in Portugal that operates a fact-checking agency known as Polígrafo and conducts special fact-checking reports. The prompts also delineated the content to be summarized automatically, bearing in mind that in a functional system, there would be a filter from the datasets to be consulted with data from recommended fact-checking agencies. In this instance, each prompt specified a different text from a fact-checking agency.

Additionally, the prompt indicated that those texts should be summarized according to certain special characteristics of each character. Different LLMs were used with a text limit imposed. This phase is critical as it defined features and personality traits in our prompts, such as "respond with the seriousness of a traditional journalist," "respond with the grumpiness of a well-informed cat," or "respond with the kindness of a beloved pet." In the subsequent stage, the summarized text automatically generates audio or text, depending on each proposed solution. Simultaneously, in the next step, the video is produced according to the necessary simulation.

In total, we presented seven solutions for experimentation, which were displayed to stakeholders in the sequence illustrated in Table 23. The characters were developed to showcase a range of approaches, from traditional to disruptive. These included well-established journalists in Portugal represented through

deepfake technology, to a grumpy cat, which was a solution suggested by participants in our previous workshop. An innovative aspect of our methodology is its emulation of what is studied in inoculation theory—namely, incorporating elements of the "virus" of misinformation into the "antidote" we propose with fact-checking. This is crucial not only for the information itself but also helps users of this solution understand how to validate false information they may encounter on the internet. To minimize the investigator bias, all solutions were presented once initially, and only during a second round were stakeholders directed to respond to each of the questionnaires in turn. Apart from the AIDA probe, which was developed by Globo specifically for this project, all other probes were created by the researcher of this project using the technologies described below.

Table 23. Technological Probes presented at the Educage Workshop

Solution	Description	Technologies used
Sônia Rebelo	Disclosure of Fact-checking using Full Fact data and Deepfake character. The character's name was created by us.	Script created with BARD API. Deepfake video created using Synthesia API.
AIDA	Disclosure of Fact-checking using Full Fact data and 3D character in Portuguese language from Globo.	Script created with BARD API. Video created using a pre-defined prototype already built by Globo's innovation team for AIDA, which has already been explained in another section of the thesis.
Dina Aguiar	Disclosure of Fact-checking using Full Fact data and deepfake from a Portuguese journalist known to the population..	Script created with Chat GPT API. Deepfake video was created using the free AKOOL API tool available for demonstration purposes.
Melinda (female cat)	Disclosure of fact-checking using data from Full Fact and a 3D character of a friendly cat. The character's name was created by us.	Script created with Chat GPT API. Text-to-speech voice created with D-ID API. Cat cartoon created with Microsoft DALL- E 3 API. Cat post-production lip sync with Davinci Resolve.
Bigodão (male cat)	Disclosure of fact-checking using data from Full Fact and a 3D character of a friendly cat. The character's name was created by us.	Script created with Chat GPT API. Text-to-speech voice created with D-ID API. Cat cartoon created with Microsoft DALL- E 3 API. Cat post-production lip sync with Davinci Resolve.
Text and audio only	Disclosure of fact-checking using data from Full Fact and text to speech only.	Script created with Chat GPT API. Deepfake video created using Synthesia API.
SIC Polígrafo	Report shown on the Sic television channel in Portugal with data from Polígrafo, a fact-checking agency belonging to the same group.	–

* The videos used, as well as the demonstration of their creation process, can be viewed on Annex B.

As for the APIs chosen, the BARD API is a large language model (LLM) that generates coherent, contextually relevant text. It operates on a vast corpus of data,

enabling it to understand and produce language that closely mimics human writing or speaking styles (Rahaman et al. 2023b). Using a combination of facial recognition and generative adversarial networks (GANs), Synthesia can superimpose new facial expressions and lip movements onto source faces in videos, effectively creating content that appears to be authentically spoken by the depicted individual (Kulkarni et al. 2023). Like the Synthesia API, the AKOOL API utilizes deep learning techniques for creating visual deepfakes but focuses on generating full body deepfake content (Cho et al. 2023). It can animate entire scenes with high realism, including body movements and interactions, which are synthesized based on input data and pre-existing video content. Chat GPT API is built on the GPT (Generative Pre-trained Transformer) architecture, a large language model known for its ability to generate human-like text based on the input it receives (Aydın and Karaarslan 2023).

D-ID API utilizes advanced neural networks that synthesizes speech that sounds natural and can be customized in terms of accent, pitch, and emotion ("D-ID" 2023). This is particularly useful in multimedia content where the voice needs to match the visual content's tone and context, ensuring a cohesive and credible auditory experience for viewers. Microsoft DALL-E 3 API is based on the DALL-E model, which itself is a variant of the GPT series adapted for image creation and generates images from textual descriptions. It uses a modified transformer model trained on a diverse dataset of images and text, enabling it to produce highly detailed and contextually accurate visuals (Hopkins 2024). Davinci Resolve is a comprehensive video editing suite renowned for its advanced post-production capabilities, which include color correction, audio post-production, and visual effects. In our research, Davinci Resolve was specifically utilized for lip-syncing features, since the API for video couldn't animate the cats, since they were only available for creating deepfakes within human faces (Wei)

Each technological setup was selected to explore distinct capabilities and integration challenges of LLMs in varying contexts. The configurations were specifically chosen to demonstrate the potential of LLMs in enhancing user engagement, crucial for an effective fact-checking system within a digital platform. This thorough investigation into each technology's specific role provided comprehensive insights into their suitability and adaptability for advancing media

production and verification processes in an academic setting. It is worth noting that the probes were displayed as a search field within the informational Moodle portal presented at the Educage project event. A non-functional prototype (Figure 44) demonstrated to the caregivers how this fact-checking search engine would be integrated into the website. Subsequently, each probe was showcased within this prototype using a projector to facilitate the continuation of the experiments.



Figure 44. High-fidelity non-functional prototype with probes for EducAge site

Considering that the purpose of our study was to co-design the social feasibility of this type of solution, rather than user tests of the technological functionalities of this solution in operation, the prototype only simulated the experiences and reactions of this type of solution within the perspectives presented in previous research with values, norms and general technical requirements. A live non-function prototype of the potential application presented can be reached by entering the link:

<https://cloud.justinmind.com/usernote/prototype/4a0e64e79db2c0d18864d6ea93d039e0d0e94af08a5c57b48b94a4532e6a2d1c>

5.4.2. Survey Implementation

The questionnaire was printed considering the participants not having individual computers in the room. The design principles proposed by Kitchenham and Pfleeger (2002) guided its construction, focusing on: a) question phrasing – ensuring the technical terminology used was appropriate for the respondents' level of understanding; quantity of questions – a balanced number of items to prevent respondent fatigue or overload; c) response options – incorporating Likert-scale and illustrative symbols for a comprehensive understanding (Figure 45); d) guidance for respondents – clear instructions enabling participants to complete the questionnaire independently. These principles ensured that the survey effectively captured the participants' feedback and provided valuable data for analysis.

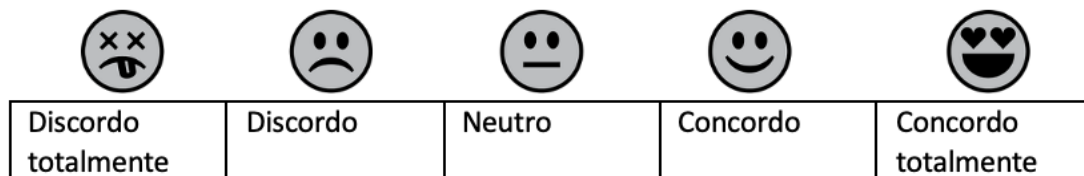


Figure 45. Questionnaire layout with response scale symbols display

The questionnaire is completely accessible in Appendix H. The values most significantly exemplified by the participants in the previous workshops that have been summarized in thematic codes will be used in the questions from the questionnaire, and are Trust, Engagement, Fear, Information, Promptness, Inclusion, Aesthetic, Equality, and Novelty. We chose to omit the transparency category from the questionnaire, opting not to delve into the technical workings of AI with our audience in this phase. We believed that providing a less detailed explanation would better simulate real-life scenarios, where individuals often struggle to grasp concepts like deepfake and discern whether a person's speech is genuine. This lack of understanding can significantly impact acceptance levels, mirroring the complexities of real-world interactions. The questionnaire was prepared in Portuguese, reflecting the nationality of the participants.

- 1) Trust to capture participants' perception of the solutions in terms of these requirements of feeling protected and trusting the tool.

- 2) Engagement, that measures the degree to which participants feel connected and familiar with the content or technology being evaluated. Considering this category to be very relevant in terms of misinformation, we have adopted here the engagement also related to the anger and fear that characters could generate.
- 3) Information, that assesses how much participants learn.
- 4) Promptness, from the speed at which it is delivered to the user.
- 5) Inclusion evaluates how well the content resonates with the audience and includes diverse perspectives or experiences.
- 6) Aesthetic, that looks at the character appearance and originality of the content or technology, gauging how visually and conceptually fresh it appears.
- 7) Equality means here appreciating diverse forms, features, and expressions of beauty without privileging one over another based on conventional standards.
- 8) Novelty, being innovative and breaking away from the ordinary or traditional.

The responses symbols are on a Likert scale, likely ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). The data shows responses from the 12 participants (PQ1-PQ12) for the solutions Sônia Rebelo, Dina Aguiar, Melinda, Bigodão, text and audio only, and Polígrafo da SIC. The provided descriptive statistics (Table 24) summarize the age distribution of a sample consisting of 2 male and 10 female participants. The sample reflects the caregiving population, which is statistically dominated by women. The female caregivers show a younger average age of 57.8 years, with a median age slightly above at 58.5 years, suggesting a younger central tendency and a left-skewed age distribution. The age sample also reflects the caregiver's population, which is statistically dominated by women in their 60s, according to data presented elsewhere in the survey.

Table 24. Descriptive Statistics

	SEX	id	AGE
N	male	2	
	female	10	
Mean	male		66.5
	female		57.8
Standard Deviantion	male		3.54
	female		16.9

This chapter delves into the analysis of questionnaire responses gathered during the collaborative workshop aimed at developing a fact-checking chatbot for elderly caregivers. By examining participants' feedback across specific themes — Protection, Trust, Engagement, Familiarity, Frustration and Frightening; and additional questions on Information, Promptness, Resonance, Inclusion, Character Appearance and Novelty —this chapter seeks to uncover insights into the design and efficacy of digital solutions in combating health misinformation.

5.4.2.1 Trust

The majority of participants (Figure 46) rated high protection levels in the information provided by the solutions. However, there was a noticeable request for more transparent sourcing of information and an easy way to verify facts, suggesting a need for embedded credibility signals within the chatbot's interface. Sônia Rebelo and Polígrafo da SIC received the highest overall scores for making participants feel protected, with several 5s, indicating a strong sense of security. These findings underscore the importance of trust and transparency in the design of fact-checking solutions. Bigodão seem to inspire moderate confidence, with most scores being 3s and 4s, but did receive a 1 from participant PQ6, indicating a significant variance in perception. The text and audio only solution also appears to make participants feel quite secure, with multiple 5s, especially from PQ2, PQ6, PQ8, and PQ12.

1. Sinto-me protegido(a) ao receber informações sobre saúde desta solução (I feel protected by receiving health information from this solution).

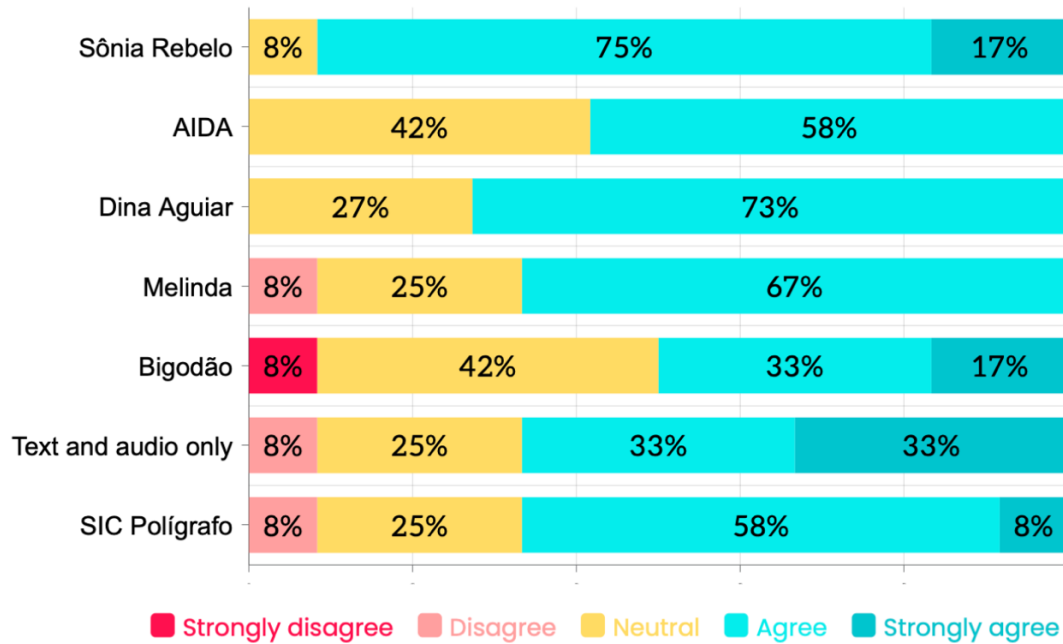


Figure 46. Grades assigned to feeling protected from each participant

Trust (Figure 47) is foundational to the user's acceptance and continued use of a digital tool, especially in sensitive areas like health information. The desire for transparency and fact verification indicates that users are not just passive consumers but active participants who value the ability to cross-check and understand the origin of the information they are presented with. Sônia Rebelo maintains a consistently high level of trust across participants, with the majority scoring 4. Dina Aguiar also appears reliable, with several 4s and a few 5s, indicating high trust. The text and audio only solution again scores highly, suggesting that participants trust the information provided in this format. Bigodão shows some variance, with scores as low as 1 from PQ6 but as high as 5 from PQ8. The Polígrafo da SIC has the widest range of responses, indicating a mixed level of trust.

2. Eu confio nas informações fornecidas por esta solução sobre saúde (I trust the health information provided by this solution)

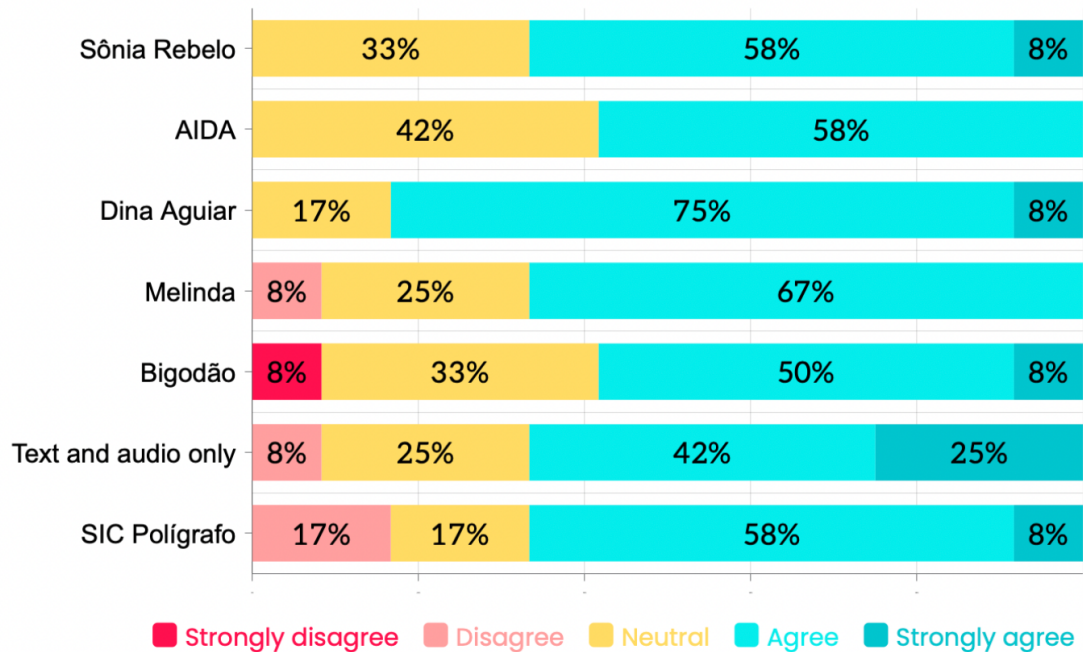


Figure 47. Grades assigned to solutions from each participant in the 'Trust' category

There is a general trend of high scores for feeling protected and trusting the information provided, suggesting that the solutions are well-received overall. Traditional formats like text and audio only are among the most trusted, which may reflect a comfort with familiar formats over more novel ones like avatars and deepfakes. The variance in responses for Bigodão and Polígrafo da SIC suggests that personal preferences and perhaps the presentation style of these solutions significantly impact individual trust and security feelings.

Regarding the implications, participants value both the feeling of security and the trustworthiness of information, which are key factors for the adoption and effectiveness of health information solutions. A broader audience may more readily trust solutions that present information in a straightforward, familiar format. The variance in responses for certain solutions suggests the need for customizable options or tailoring content to different user.

5.4.2.2 Engagement

Responses indicated a preference for a user interface that mimics familiar social media platforms, with intuitive navigation and terminology that is easy to understand for non-technical users. The participant feedback on the health information solutions in terms of involvement (Figure 48) reveals a clear preference pattern. This feedback highlights the critical role of user engagement in the effectiveness of health information solutions. Sônia Rebelo and Dina Aguiar stand out, consistently receiving scores of 4 and above, indicating that these solutions successfully captivated the participants' interest in health topics. Melinda, while highly engaging for certain participants, received a mix of scores, suggesting its impact varies across the user base. Bigodão achieved moderate involvement levels, indicating a solid, if not stellar, performance in drawing users in. The text and audio-only solution, meanwhile, saw polarized reactions, resonating strongly with some users but less so with others, which might reflect personal preferences in learning modality. Polígrafo da SIC maintained a consistent level of engagement, with most users finding it to be a reliable and motivating source of health information.

3. As respostas desta solução foram envolventes e me incentivaram sobre saúde (The answers to this solution were engaging and encouraged me about health)

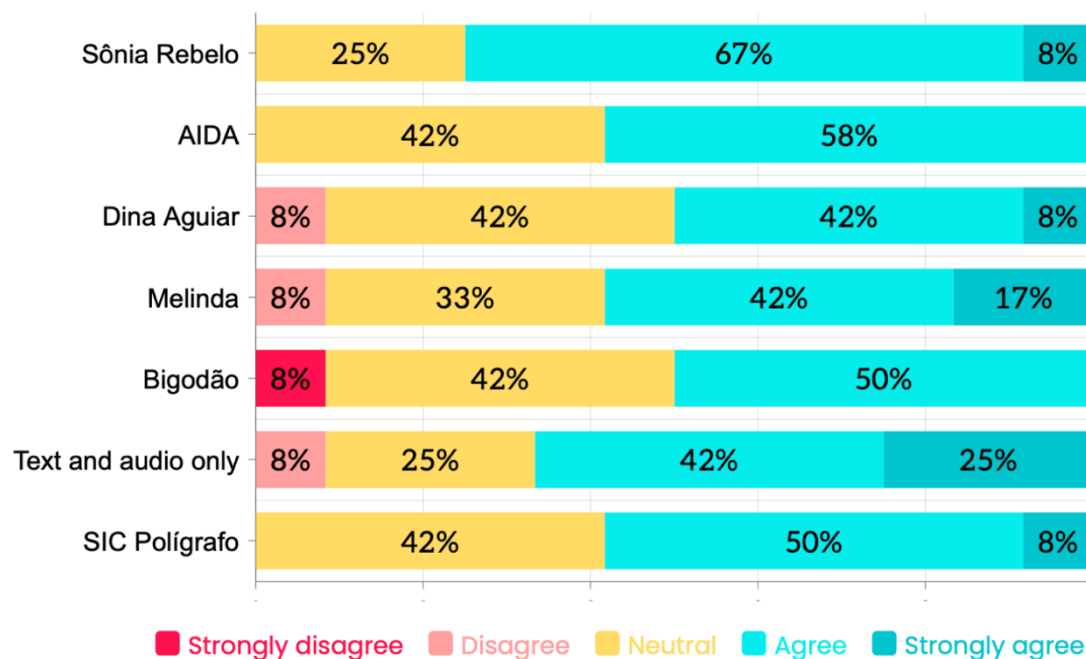


Figure 48. Grades assigned to solutions from each participant in the 'Engagement' category

Familiarity breeds comfort (Figure 49), which in turn can increase the usability and acceptance of new technologies among older populations. Designing with an eye toward what is already well-understood by the target demographic can decrease the learning curve and enhance the overall user experience. When it comes to the familiarity of language used in the health information solutions, the data suggests that Sônia Rebelo effectively communicated in a manner that users found very approachable, as evidenced by several top marks.

Dina Aguiar's use of language was consistently familiar to participants, aligning well with their understanding. Melinda showed a range of responses, indicating that while the solution was perfectly clear to some, it may not have universally met all users' preferences for familiar terminology. Bigodão's language was mostly understood, although there's room to enhance its relatability. The text and audio-only solution again displayed a split in user responses, and Polígrafo da SIC was seen as largely successful in using language that the majority of participants felt comfortable with, aligning well with their expectations for clarity and ease of comprehension.

4. Esta solução usa uma linguagem que me é familiar para aprender sobre saúde (This solution uses a language that is familiar to me for learning about health)

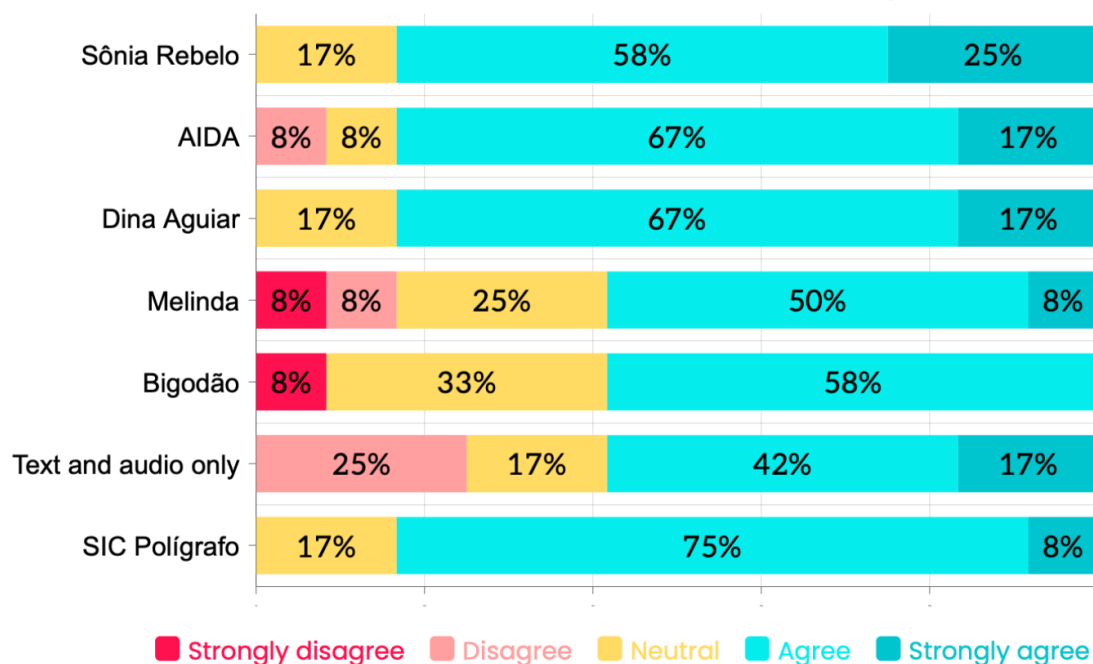


Figure 49. Grades assigned in the “engagement” category regarding resonance

Some users reported feeling anxious when the chatbot presented information that contradicted their pre-existing beliefs or when the language used was perceived as too authoritative or impersonal. This feedback highlights the delicate balance required in communicating fact-checked information. The scores for this section are mostly low (1s and 2s), which suggests that for most participants, the solutions did not cause significant irritation or confusion. However, there are some exceptions (Figure 50). Sônia Rebelo shows a higher level of irritation or confusion in some responses, particularly from PQ4 and PQ7, who rated it a 4.

This indicates that while some participants found the solution generally unproblematic, it may have specific aspects that could be refined to reduce any potential frustration or misunderstanding. The text and audio only solution generally scored low for irritation or confusion, but PQ8 and PQ9 rated it a 5, suggesting a strong negative reaction from these individuals.

5. As respostas desta solução me irritaram ou confundiram durante a interação (The answers in this solution irritated or confused me during the interaction)

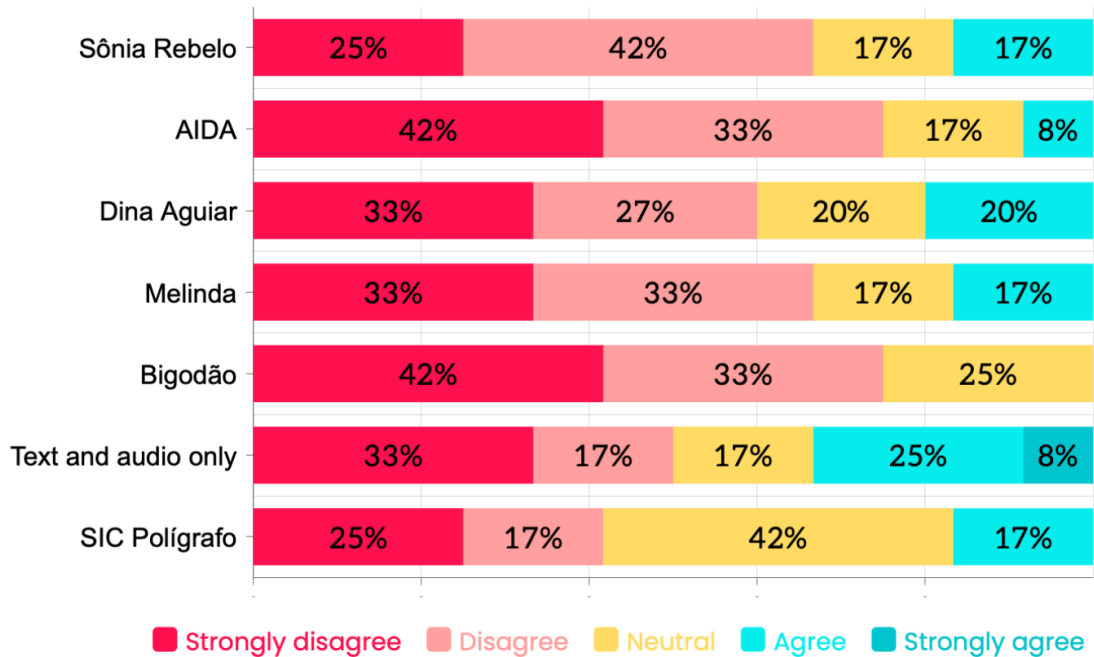


Figure 50. Grades assigned to solutions from each participant about “Frustration”

The tone, framing, and presentation style can significantly impact the user's emotional response. Personalization and careful consideration of language can mitigate negative reactions and support a more constructive engagement with challenging topics. Regarding fear (Figure 51), most scores are low (1s and 2s), indicating that the information provided typically did not provoke worry or fear about health among participants. Sônia Rebelo received a 4 from PQ4 and PQ10, indicating some instances where the information may have caused concern.

6. As informações fornecidas por esta solução geraram preocupação ou medo sobre minha saúde (The information provided by this solution generated concern or fear about my health)

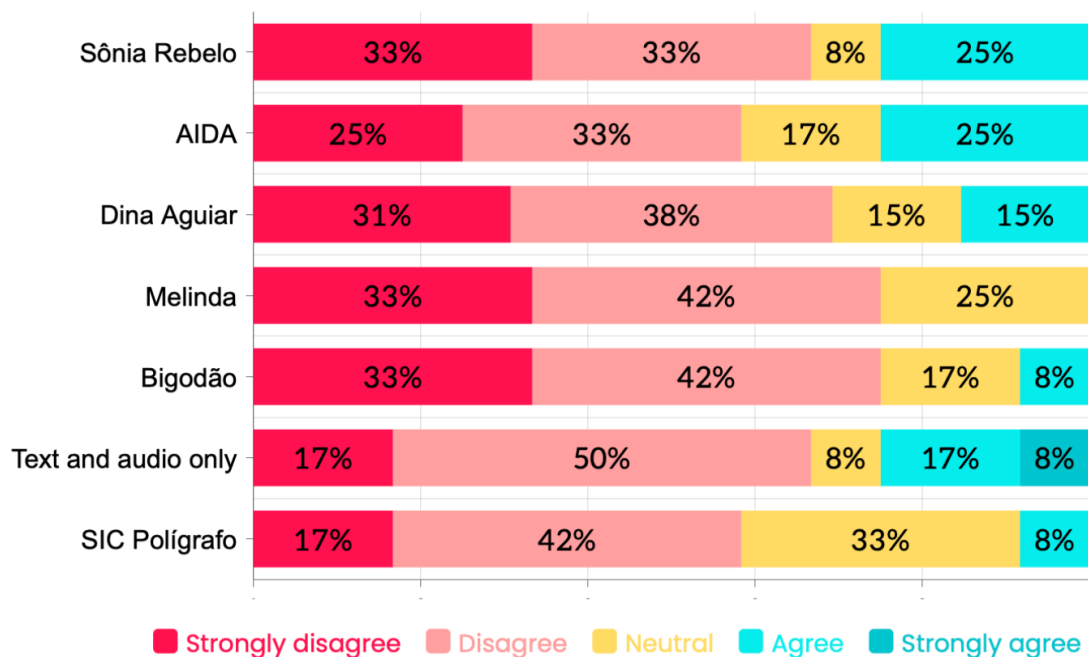


Figure 51. Grades assigned to solutions from each participant about “Frightening”

The data reflects a positive reception in terms of emotional response, with most solutions not causing significant irritation, confusion, worry, or fear. Instances where higher scores are present suggest areas for potential improvement. It may be beneficial to examine the content or delivery method for those specific solutions to identify what might be causing the stronger reactions. The outliers where higher scores were given could also be due to individual differences among participants, such as personal health concerns or sensitivity to the information presented.

Considering these, we would conduct a deeper qualitative analysis to understand the context of the higher scores for irritation, confusion, worry, or fear. Understanding the "why" behind these responses can guide improvements in the solutions' design and content delivery. The focus group, then, should consider personalizing the user experience to accommodate a wide range of emotional responses and sensitivities to health information. In summary, while the overall emotional response to the solutions is positive, attention should be given to instances of higher scores indicating negative reactions. These insights can inform

targeted refinements to enhance the user experience and ensure the solutions are supportive and not distressing for the intended audience.

5.4.2.3 Additional Themes of engagement and resonance

5.4.2.3.1 Information and Promptness

Analysis of "Interesting Health Information" (Figure 51), Sônia Rebelo scores consistently high, indicating that the majority of participants find the information presented by this solution to be interesting, which is key for engagement and continued use. Dina Aguiar has a strong positive reception, especially from PQ8 and PQ12, who rated it a 5. Melinda also performs well, with many 4s and some 5s, indicating that the information presented is engaging to most participants. Bigodão and text and audio only solutions receive a mix of 3s and 4s, which could imply that while the information is satisfactory, it may not be as captivating as that from other solutions. Polígrafo da SIC receives varied responses, with a mix of 3s, 4s, and a couple of 5s, which might indicate variability in how interesting different users find the health information presented.

7. Esta solução apresenta informações interessantes sobre saúde (this solution provides interesting health information)

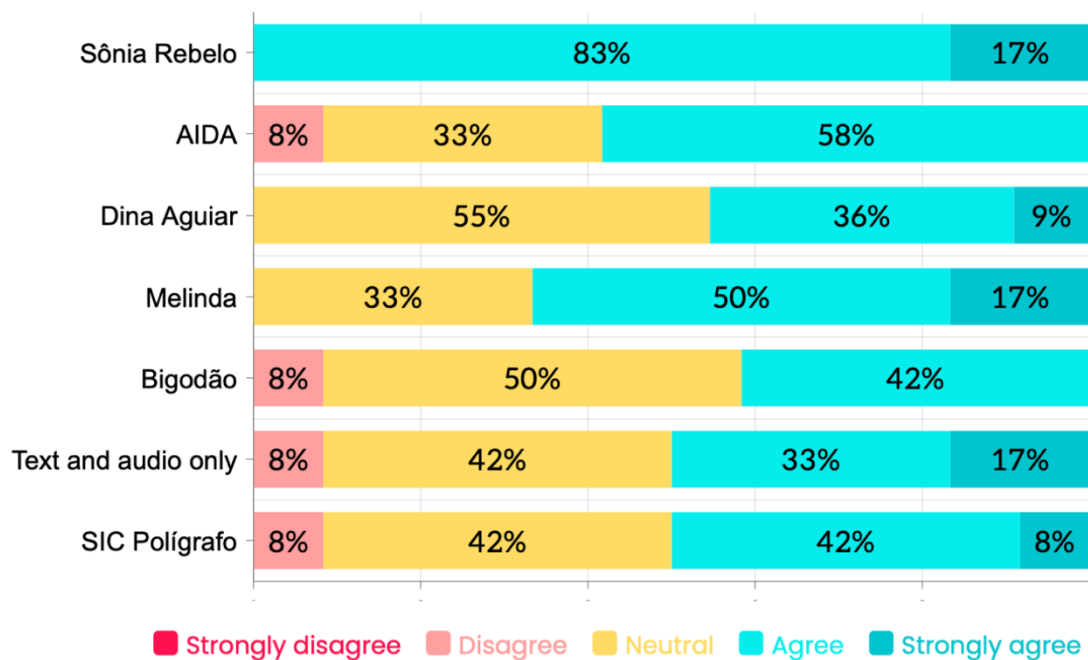


Figure 52. Grades assigned to solutions from each participant in the "Information" category

Related to "Importance of Immediate Response" (Figure 53), Sônia Rebelo and Dina Aguiar rate highly on the importance of immediate responses, which suggests that the promptness of these solutions in providing information is valued by the users. Text and audio only solutions receive a mix of 3s, 4s, and some 5s, indicating that while immediate response is appreciated, there may be some tolerance for slight delays. Melinda shows consistency in scoring 4s, implying that users find immediate responses important, but not necessarily critical. Bigodão and Polígrafo da SIC present a consistent pattern of 3s and 4s, which could mean that immediate responses are important and would be crucial to a solution such as Polígrafo, that requires some time to be prepared, since it is not automatized.

8. É importante receber essa resposta desta solução imediata à pergunta que fiz no verificador virtual (It's important to receive that answer from this immediate solution to the question I asked in the virtual checker)

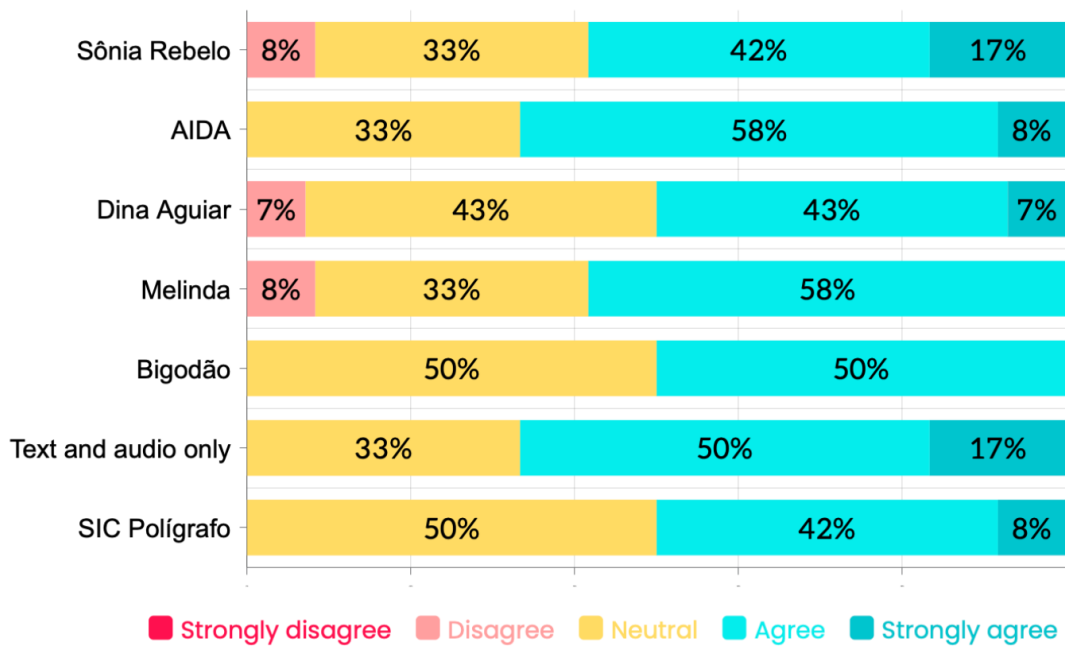


Figure 53. Grades assigned to solutions from each participant in the "Promptness" category

Across the board, interesting health information is rated quite positively, which is crucial for the effectiveness of health communication solutions. There is a general tendency for users to value immediacy in responses, which highlights the need for the solutions to be responsive in real-time or near real-time to maintain user

engagement and trust, specially under infodemic scenarios. For solutions like Bigodão, where the scores indicate room for improvement, further investigation into the content could reveal what users might find more engaging or interesting. The importance of immediate responses underlines the necessity of having robust and efficient backend processes to ensure that users get timely information, which is especially vital in the health context where decisions may be time-sensitive.

5.4.2.3.2 Equality and Inclusion

Participants valued features that solicited their input, such as feedback options and customization settings, seeing them as a way to feel more involved in the information exchange process (Figure 54). Analysis of "Motivation to Interact More About Health", Sônia Rebelo scores highly across most participants, indicating that this solution strongly motivates users to engage more with health information, which is crucial for user retention and ongoing interaction. Dina Aguiar and Polígrafo da SIC, with consistent 4s, appear to encourage interaction, signaling effective user engagement. Melinda receives a mixed response but still maintains a decent motivational influence, especially noted from PQ1 and PQ8. Bigodão and text and audio only solutions show a consistent pattern of 3s and 4s, indicating a general but not universal motivation to interact more. PQ6 consistently scores lower for several solutions, which may indicate individual differences in response to the solutions or a specific user need that is not being met.

9. O uso do verificador virtual com esta solução me incentivou a interagir mais sobre saúde (Using the virtual checker with this solution encouraged me to interact more about health)

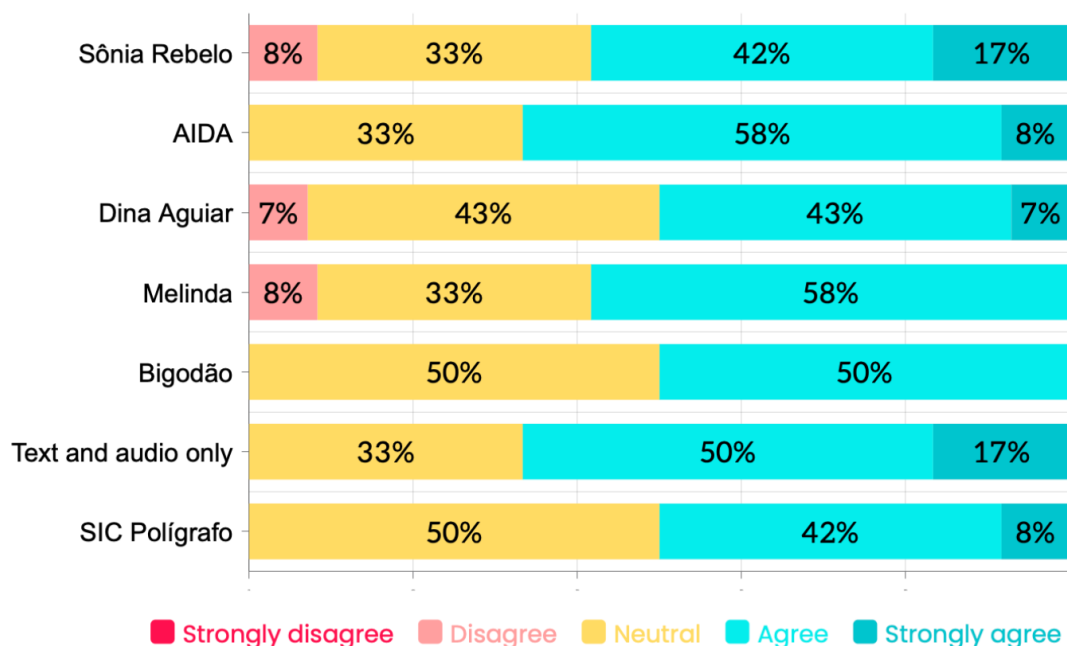


Figure 54. Grades assigned to solutions from each participant in the "Resonance" category

Active involvement and personalization features can foster a sense of ownership and control over the user experience, enhancing engagement. Inclusion, in this context, also means accommodating diverse needs and preferences, reinforcing the chatbot's role as a supportive tool for a wide range of users. Analysis of "Feeling Represented as a User" (Figure 55), Sônia Rebelo has strong scores, with a 5 from PQ11, showing that users often feel represented when using this solution, which can increase personal relevance and trust in the information provided. Dina Aguiar, Melinda, and Bigodão generally score well, suggesting they are doing a decent job at making users feel represented. The text and audio only solution has mostly 3s and 4s, with a high 5 from PQ8, reflecting that traditional formats can still be effective in user representation. Polígrafo da SIC scores consistently but has a few lower scores from PQ2 and PQ10, which could suggest certain aspects may not resonate as strongly with some users.

10. Senti-me representado(a) enquanto usuário(a) ao interagir com esta solução a me passar informações sobre saúde (I felt represented as a user when I interacted with this solution to provide me with health)

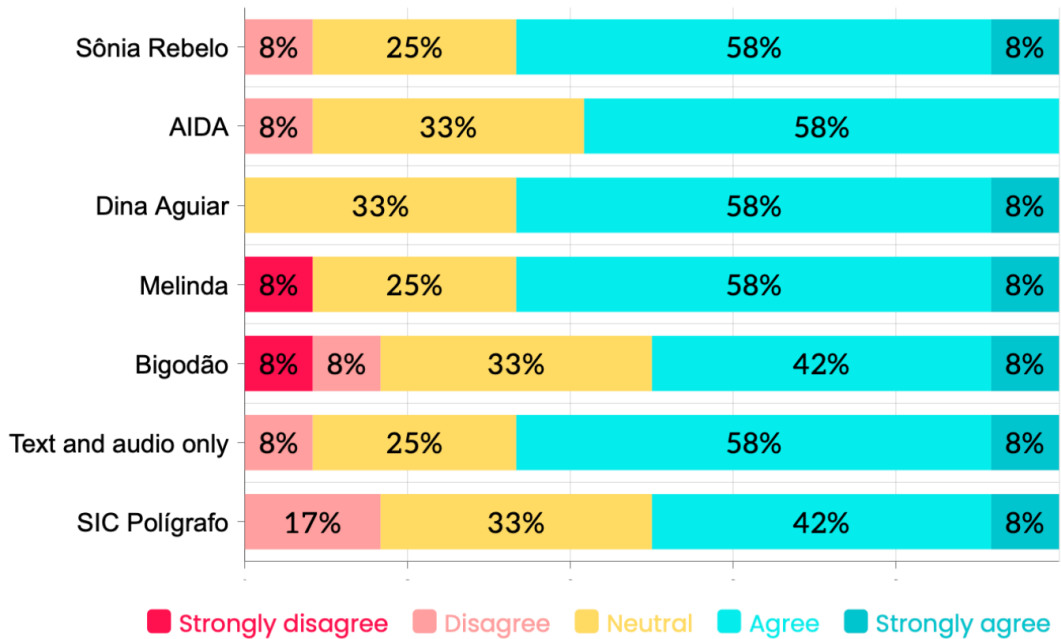


Figure 55. Grades assigned to solutions from each participant in the "Inclusion" category

The data suggests that most of the solutions are effective in motivating users to interact more and feel represented, which are key aspects of user experience and satisfaction. Solutions that score highly on these aspects, like Sônia Rebelo and Dina Aguiar, could be leveraged as models for further development of similar health information tools. Lower scores in certain areas could indicate a need for personalization or more user-centric design features to improve representation and motivation for all users.

5.4.2.3.3 Aesthetic and Novelty

Aesthetic preferences varied widely among participants, but there was a general consensus that the chatbot should have a friendly and approachable appearance, with some suggesting that customizable avatars could enhance the appeal. The latest data sets provided offer insights into how participants perceive the visual presentation, the impact of the solution's appearance on information comprehension, and how the solution's personality influences user comfort. Let's analyze the responses for each statement.

Regarding visual authenticity and trustworthiness (Figure 56) for learning about health, Sônia Rebelo generally is perceived as authentic and trustworthy, indicated by the high scores, with a notable score of 5 from PQ12. Dina Aguiar and Polígrafo da SIC are consistently rated well, suggesting a strong visual trustworthiness. Melinda and Bigodão have high scores from some, indicating visual appeal and trust, but lower scores from others, like PQ6, suggest not all users feel the same. The text and audio-only solution is generally well-received, but there may be room for improvement in visual aspects for certain users, like PQ1.

11. A maneira como esta solução se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde (The way this solution looks visually seemed authentic and reliable for learning about health)

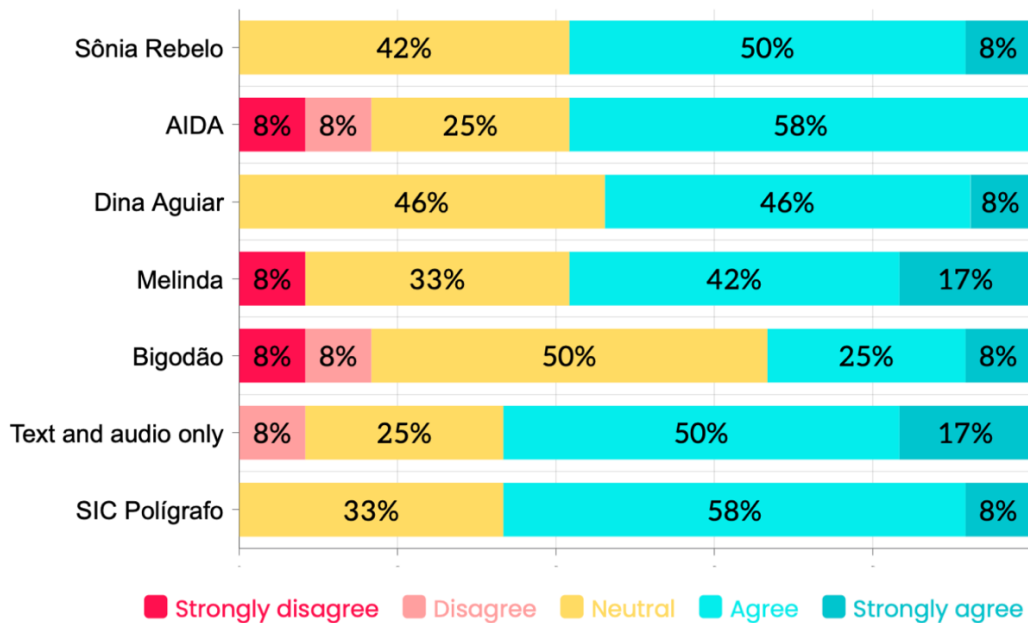


Figure 56. Grades assigned to solutions from each participant in "Equality"

The visual design of a chatbot (Figure 57) can significantly affect its approachability and the user's willingness to interact. While beauty is subjective, the option to customize appearance can empower users, allowing them to tailor the chatbot to their personal preferences, which could improve satisfaction and long-term engagement. Impact of Solution's Appearance on Health Information Comprehension also presents Sônia Rebelo scores that reflect its appearance contributes positively to making health information more interesting and comprehensible, especially noted with a score of 5 from PQ7 and PQ11. Dina Aguiar has strong positive feedback, with several 5s, indicating a significant

contribution to comprehension and interest. Melinda, Bigodão, and the text and audio only solution have good scores, suggesting their appearance aids in understanding, with room for minor improvements. Polígrafo da SIC receives varied scores but is generally positive, suggesting its appearance helps in understanding health information.

12. O aspeto desta solução contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis (The look of this solution has helped to make health information more interesting and understandable)

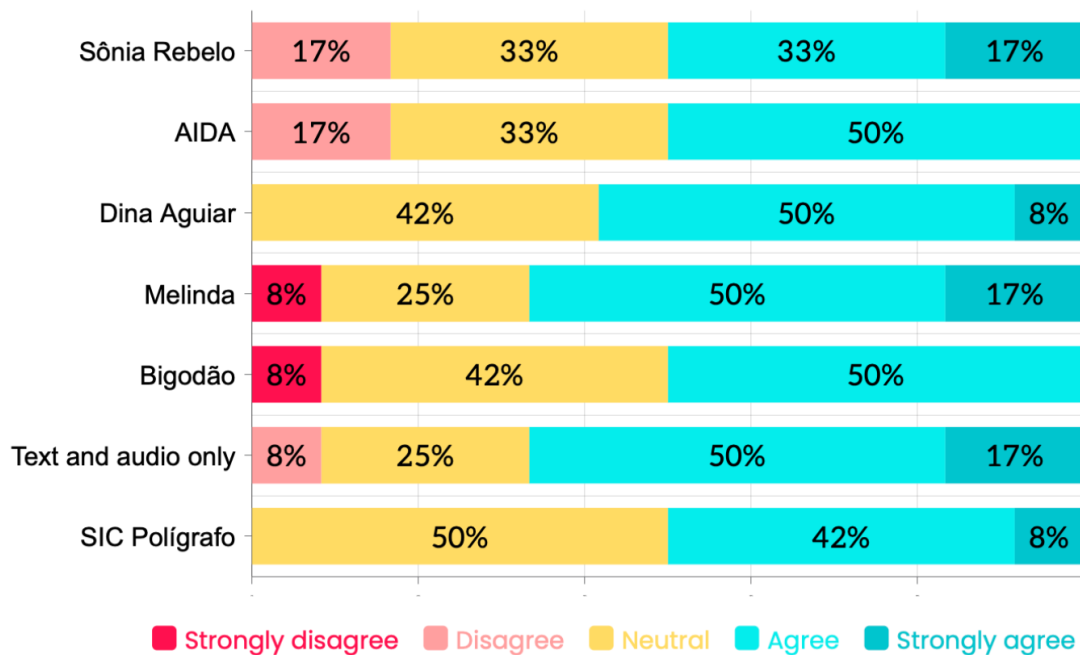


Figure 57. Grades assigned to solutions from each participant in the "Aesthetics" category

According to the criterion of "personality", an item highlighted in the previous workshop, Influence of solution's user comfort (Figure 58), Sônia Rebelo has mostly 3s and 4s, which suggests its personality is relatively effective in making users comfortable, but there may be potential to enhance this aspect. Polígrafo da SIC receive a good range of scores, indicating that certain personality traits resonate well with users. Dina Aguiar receives consistently strong scores, suggesting that its personality greatly aids user comfort. Melinda and Bigodão have a high score of 5 from several participants, reflecting a personality that significantly aids in user comfort. The text and audio only solution is perceived as comfortable and

accessible, with consistent scores across the board, which may reflect well on its user-friendliness and approachability.

13. A personalidade desta solução (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde (The personality of this solution (e.g. friendly, serious, approachable) helped me feel more at ease when searching for health information)

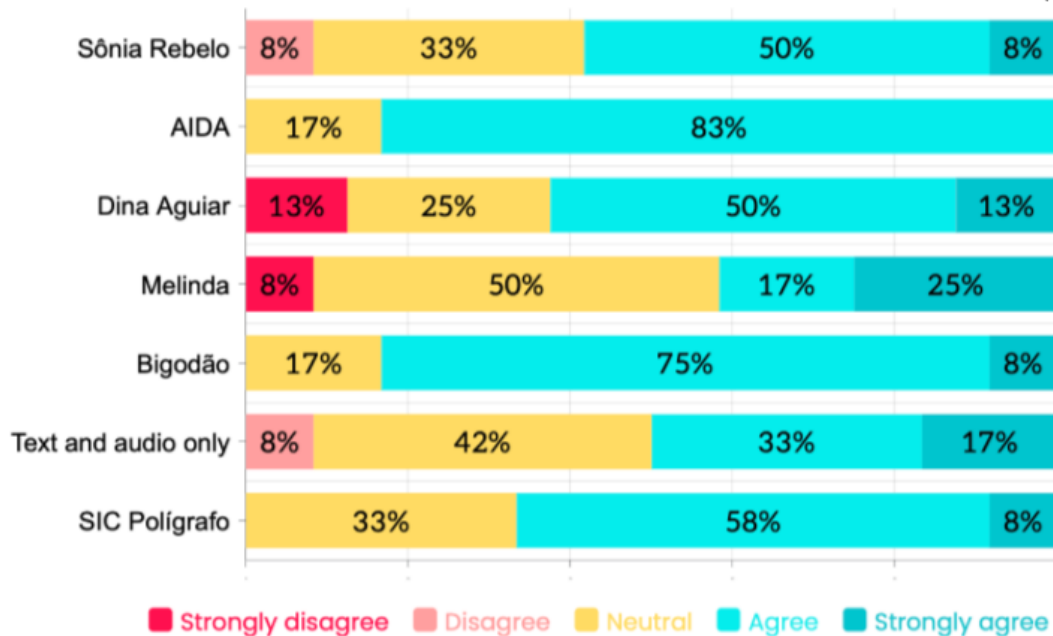


Figure 58. Grades assigned to solutions from each participant in the 'Novelty' category

According to these scenarios, visual authenticity and trustworthiness are important factors in users' learning experiences, and solutions that score well in this aspect are likely to be more effective. The appearance of the solution significantly impacts the interest and comprehension of health information, suggesting that design elements should be carefully considered to enhance these aspects. The personality of the solution, including traits like friendliness and accessibility, plays a crucial role in how comfortable users feel when seeking health information.

5.4.2.4 Preferred solutions

The final chart (Figure 59) presents a compilation of user preferences across various health information solutions, with participants asked to rank these from their

most to least favorite. In an assessment of overall satisfaction, Sônia Rebelo and Dina Aguiar emerged as the top choices, each accumulating a total of 48 points across all participants. This outcome suggests these two solutions were equally favored among the users, resonating well with their needs and preferences.

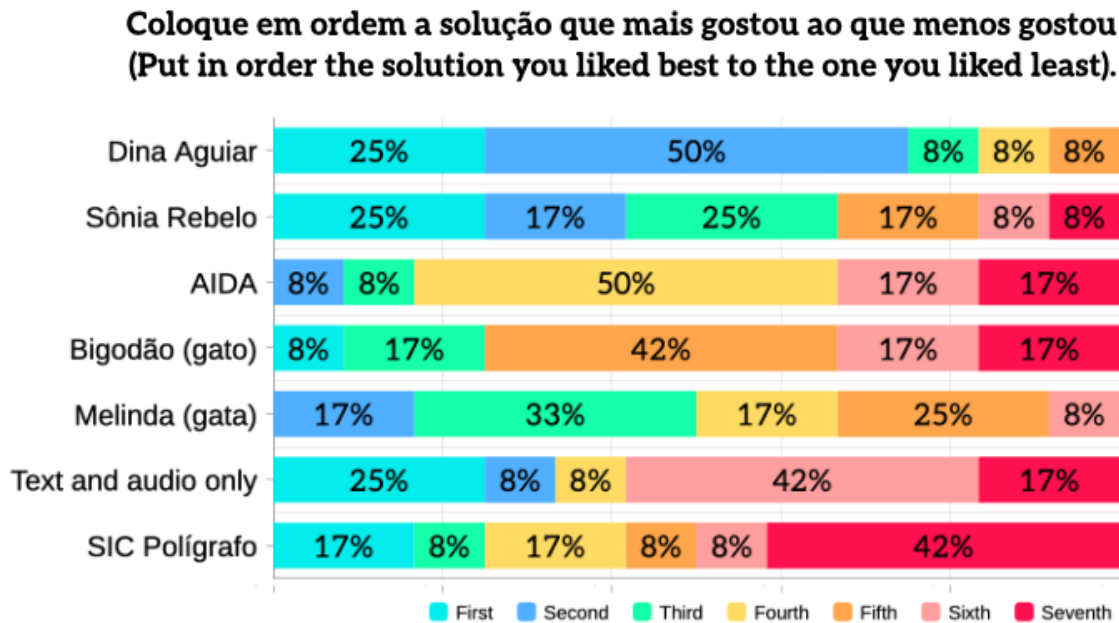


Figure 59. Chart presenting character ranking

Sônia Rebelo shows a diverse range of rankings but notably secures the top spot (6 points) from PQ1, PQ6, and PQ9, indicating very strong preferences from certain individuals. Meanwhile, Dina Aguiar appears to be a consistent favorite, achieving the highest ranking (6 points) notably from PQ4, PQ7, PQ9, and PQ10, showing a broad appeal among the participants.

Melinda, with a total of 39 points, sits in the middle tier, indicating a generally favorable perception, although slightly less preferred when compared to the leaders. It scored quite high with PQ2 and PQ11, securing 5 points, but didn't manage to consistently hit the top spot. Text and audio only, while not the most favored, still garnered a significant total of 28 points, including top rankings (6 points) from PQ3, PQ5, and PQ12, suggesting that when it was the preferred choice, it was strongly so. Bigodão, with a total of 22 points, and SIC Polygraph, with 21 points, were less preferred overall, with SIC Polygraph notably not receiving any points from half of the participants. However, it's important to note that SIC

Polygraph was the top choice (6 points) from PQ2 and PQ11, indicating that for a specific subset of users, it was the preferred solution.

In conclusion, while Sônia Rebelo and Dina Aguiar stood out as the most appreciated solutions for health information, indicating their effectiveness in user engagement and content delivery, there were varied preferences across the board. This diversity in user rankings underscores the importance of offering a range of solutions to cater to the different needs and tastes within any user base.

5.4.2.5 Insights from the Survey

The insights garnered from the responses of participants in survey, specifically aimed at the elderly, are invaluable for guiding the development of the proposed technological tool. Trust and security emerged as prominent aspects. Solutions such as Sônia Rebelo and Dina Aguiar received high trust ratings, indicating that the use of familiar figures in virtual representation can make participants feel protected and secure. The data reveals a significant demand for greater transparency regarding the origin of information and easy methods for fact verification, suggesting the need for embedded credibility signals within the fact-checker interface.

Regarding engagement and familiarity, participants preferred interfaces that mimicked familiar social media platforms, featuring intuitive navigation and easily understandable terminology. This indicates that familiarity can enhance the usability and acceptance of new technologies. The promptness and quality of information were highly valued, a critical point justifying the confirmation of using these AI technologies. There was also a notable preference for solutions involving real people, even in their deep learning versions. The diversity of preferences highlighted the importance of offering customizable options to meet the varied needs of users, including different levels of digital literacy and accessibility requirements, as well as the use of audio-only solutions for some participants.

Aesthetics and novelty played a significant role in user engagement. There was a general consensus that the fact-checker should have a friendly and approachable appearance, with the customization of avatars potentially enhancing attractiveness. The visual presentation of the solutions significantly impacted users' trust and

comprehension of the information, with solutions like Sônia Rebelo and Dina Aguiar being perceived as visually reliable.

Finally, the use of avatars and deepfake technologies offered benefits such as speed and accessibility. However, concerns about the authenticity and trustworthiness of avatars and deepfakes were noted, with a clear preference for familiar audio formats. It is also interesting to note that more playful 3D solutions were not outright rejected, though they were not the top choices in the overall ranking. The next phase, in which we carried out a more in-depth survey in focus group format on these values, shed more light on these evaluations.

5.4.3 Focus Group Analysis

The previous quantitative analysis described their opinions with predetermined values and principles. We intend to delve deeper into understanding the reasoning behind these evaluations through qualitative methods. By interviewing users with focus groups and providing co-design exercises, we aim to gain detailed insights into how digital fact-checkers meet their expectations and whether these align with the claims made by previous phases of our research.

In summary, our study employs quantitative and qualitative methods to comprehensively understand digital avatars and automated audio, aiming to uncover patterns, values perceptions, and communication strategies within the digital fact-checking landscape. The focus group reveals a nuanced discussion on health information solutions among caregivers (Appendix O). Some participants preferred Dina Aguiar and text and audio formats, highlighting the importance of clear and direct information delivery and an AI solution for co-existence in real and virtual worlds. Some found the characters and virtual elements less appealing, suggesting a divide between those who favor a more traditional, straightforward approach and those who may appreciate innovative, interactive methods.

Concerns were raised about the relatability of virtual characters and the effectiveness of humorous approaches in serious health contexts. Participants also discussed the accessibility of digital tools for older populations and the varying levels of internet proficiency among caregivers, emphasizing the need for inclusive solutions that cater to a range of digital literacy levels. The conversation pointed

towards the potential of technology to aid in disseminating information and the desire for trustworthy, easily verifiable sources, reflecting a collective awareness of the challenges in navigating health information in the digital age.

During the focus groups, participants engaged in a rich discussion (Appendix N) about the various health information solutions they experienced. Some of their answers were translated from Portuguese here:

Table 25. Highlights on first impressions of solutions by theme

Sub-themes	Participant	Asserts Example
User experience design	PQ8	I prefer something more real than virtual... The voice and information do not present as negative or false. But a virtual image does raise more doubts for me about how the response might be.
Unique features	PQ11	If it were just to create an icon and have it in audio and text, that would be fine.
Accuracy of health-related information	PQ4	I always look for scientific sites. Moreover, that there is a reference.
Representation of diverse groups	PQ3	From the age of 60 and 70, many people do not know how to use the internet. They are left out of this solution.

The focus group also touched on the Informative potential of these health information tools to assist caregivers in their roles. PQ10 voiced the need to educate caregivers on digital tools: "I want to teach caregivers how to access this." PQ6 raised a critical change in the educational background between the elderly and their caregivers, suggesting that caregivers are now more informed and often have higher education levels than those they care for.

Also, regarding the "Information" thematic code, PQ2 shared a personal anecdote about caring for her husband, illustrating the practical need for accessible health information among caregivers, "I did not have training. However, I had to care for my husband." The group also touched on the broader implications of such tools for elderly care, with PQ9 suggesting that the country needs to be more attentive and PQ11 advocating for self-empowerment, "If we want the elderly to be treated differently, or better, we have to do something about it."

The focus group transcript offers insightful perspectives from participants with caregiving experiences, providing a deeper understanding of their needs and preferences in health information solutions.

Table 26. Highlights of caregiver experiences by theme

Sub-themes	Participant	Asserts Example
Digital Accessibility	PQ3	I worked in a nursing home for elderly women in Switzerland, in Lausanne... It was the best because the three houses were in the same place...those who could look after themselves. Furthermore, they had television. Everyone had television. Everyone had their own will. (...) And the interesting thing there is that they could stay up much later together. Here in our homes, there is no one because the elderly have to be all in bed.
Creativity and originality	PQ8	So, for five years, I transported sick people. I had two experiences with some Alzheimer's people. One, known. I used music in the car every day. And I have even recorded music. And we played guessing the singer. They said the name. Look, my strength isn't in singing. And they sang the song. Music greatly improved people with Alzheimer's, and man has already transformed.
Representation of diverse groups	PQ3	Furthermore, they don't have access to television. There I realized that they want the television. They want, for example, the soap operas. They love to watch soap operas. They want to see the programs...

These participants' observation highlights the importance of entertainment and personal choice in care facilities, suggesting that such amenities can significantly impact the well-being of the elderly. They also highlight the therapeutic potential of music and personalized interaction in Alzheimer's care, emphasizing non-traditional methods that can offer significant benefits. The negative comment also touches on the challenges in providing health information for individuals with Alzheimer's:

Table 27. Highlights on experiences with innovations

Sub-themes	Participant	Asserts Example
Availability and accessibility	PQ12	But, in a situation of Alzheimer's, sometimes even the doctors don't give good advice
Reliability and consistency	PQ4	You have to have your own knowledge. Then, realizing what is good for him, what is bad. How is our health system? Sometimes there is no doctor. Because the doctors, you know? They don't inform the patient; they don't say what should be. For example, they need a specialty. They do not have it because they don't even know the specialist doctor.

These excerpts from the focus group discussion reveal the caregivers' insights into the nuanced needs of the elderly and those with cognitive challenges. The participants advocate for respectful, informed, and personalized approaches to

care, stressing the need for accessible health information and the role of caregivers in ensuring the well-being of their charges. They underscore the diverse opinions and needs within the caregiver community regarding health information solutions. The conversation reveals a clear preference for realistic, direct, and accessible information, highlighting the importance of considering the end-user's background, preferences, and environment when developing and implementing such tools.

In the preceding analysis, we evaluated how the essential values of our technological tool were perceived by the public. It was significant to observe that, unlike participants from other phases of our workshops—who focused more on principles of journalistic ethics or broader notions of equality—this specific audience, represented by elderly caregivers, perceived distinct ethical challenges. These challenges primarily revolved around the inaccessibility of medical professionals or family members in providing accurate information. The caregivers clearly expressed a need for inclusion tailored to regional contexts and inclusive communication, which radio journalism can facilitate. Furthermore, the audiovisual medium's potential to bring people together in a home setting for shared conversations about their final days underscores a sense of belonging and inclusion. This factor of belonging and inclusion should be highlighted and leveraged in developing a technology that fosters a connection with the message to be viewed and subsequently shared on social media.

In the next chapter of the thesis, we engage in an intriguing design activity where participants, through the use of the empathy map, balance these values via a focused discussion. They consider how the character representing them should be crafted. This approach was more adapted to this audience, enabling them to choose the most suitable form of representation. We believe it is an effective way to creatively envision the use of the tool. Here, we distinguish this phase from the previous one, where we employed envisioning cards and had participants discuss value tensions. We found that such a discussion might be too academic for this audience; thus, we opted for a more practical and playful activity. Nonetheless, it still required them to decide how to balance values among themselves. The result yielded rich and heterogeneous scenarios, which will be described in the following section.

5.5.4 Empathy map brainstorming

As a result, from empathy map (Appendix L), we had several insights into the responsiveness of these solutions to target audiences. In the first phase, the questionnaires provided participants' opinions on each of the solutions presented and allowed us to rank which ones are most relevant within each of the pre-defined categories of the newness of attention, protection and trust, involvement and familiarity, anger, and fear, as well as to understand the receptiveness of identification and acceptance of each solution. We found that there was no great differentiation or repulsion towards the traditional fact-checking initiatives and the automated ones, nor was there any great dislike of the 3D animation solutions with pets.

During the focus group, the participants gave more details about these perceptions (Appendix M), and in their responses, we realized that there was dissatisfaction with the reception given by the health institutions. One interesting opinion came from the highlighting of how the elderly have a place to meet each other on television in the lounges of nursing homes and that these new technologies should consider the welcoming factor as something to be emphasized. In the empathy map workshop, three distinct groups (Figure 60) developed unique characters, each reflecting varied nuances of life experiences in Macedo de Cavaleiros.

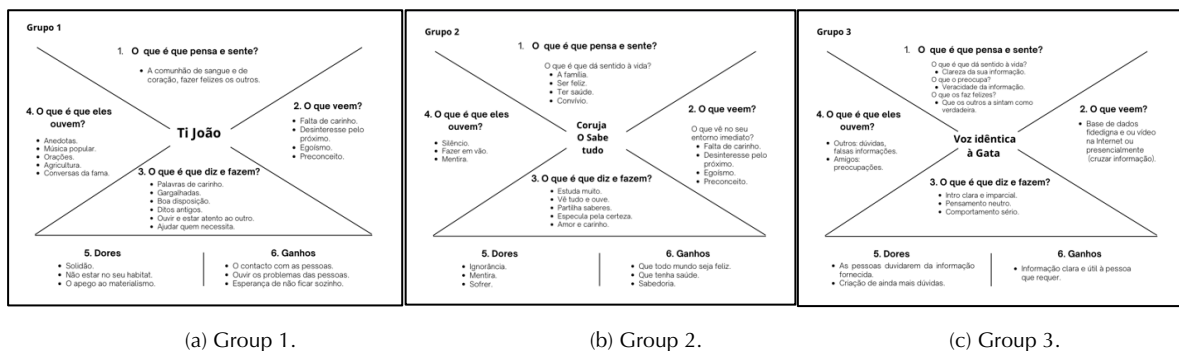


Figure 60. Three papersheet work resulting from workshop subgroup's exercise

The first group, consisting of PQ4, PQ12, PQ10, and PQ6, created the character Ti João, inspired by a local radio journalist. Ti João finds meaning in life

by bringing joy to others, standing out for his ability to listen and meet the needs of others. He lives in a world he perceives as lacking affection and solidarity, marked by disinterest and prejudice. Expressing himself through words of affection and laughter, Ti João is concerned with loneliness and materialism, longing for deep and genuine human connections. From the audio transcription of this group presentation from PQ6, we could highlight in Table 28:

Table 28. Group 1 participants explain their choices

Sub-themes	Participant	Asserts Example
Reliability and consistency	PQ6	So, we chose a character inspired by radio announcers like Rui Costa and Uncle João from Bragança. They can reach out further in terms of communication and speaking with them.
Participation and involvement	PQ6	There is a difference compared to the characters of this project: They can listen. That could be a tip to leave that it could listen to people's problems. Because throughout all the things we were adding, listening was a fundamental aspect.

The second group (PQ8, PQ2, PQ3, PQ9) introduced the Owl, a character symbolizing wisdom and perception. The Owl values family, happiness, and health but is tormented by inequality and loneliness. Visually, this character may reflect the wisdom traditionally attributed to owls, with a keen observation of its surroundings, notably similar to the scenario perceived by Ti João. Studious and reflective, the Owl seeks to convey knowledge and love while fighting against ignorance and aspiring to a world where everyone is happy and wise. These are the highlights (Table 29) from this group presentation:

Table 29. Group 2 participants explain their choices

Sub-themes	Participant	Asserts Example
Availability and accessibility	PQ3	So, the main thing is family. We have someone from the family to help us be happy and healthy, with coexistence, union, guilt, inequality, loneliness, abandonment, and lack of love.
Cultural sensitivity	PQ3	And now here, what do they hear? Silence. Often, we need to correct them and remain silent.
Clarity and comprehensibility	PQ3	What do they say and do? The owl. What do they allow us? She studies, sees everything, hears, shares knowledge, and speculates for certainty, love, and affection.

Finally, participants PQ1, PQ11, PQ5, and PQ7, with the third group (Table 30), focused on their character's voice without reaching a consensus on its visual appearance. The chosen voice is similar to that of a cat presented earlier, suggesting a gentle yet assertive approach. This character, valuing clarity and truthfulness of information, is deeply concerned with the authenticity and accuracy of the data shared. Their voice would be used to express neutral and impartial thoughts, concerned with the reception and impact of their information, always in pursuit of elucidating truths and dispelling doubts. Here are the highlights of this group:

Table 30. Group 3 participants explain their choices

Sub-themes	Participant	Asserts Example
Accuracy of health-related information	PQ1	We did not choose any character because we chose different characters... However, we recommend that the tone of voice or the audio must always accompany the text or character, which creates some closeness. So, the tone of voice we chose was that of Melinda, the cat, who, in times of difficulty, can help you in this passage of information. Therefore, we chose a voice with the audio so you do not have those feelings as if it were a person.
Interactivity mechanisms	PQ5	The image makes little difference. The voice, tone, or audio perhaps creates closeness and verifies what we hear.

In summary, each group created a character with a distinct personality and perspective, ranging from Ti João, who embodies the human warmth and compassion of a community journalist, to the reflective and wise Owl, and the cautious and accurate voice of the third character, all reflecting different aspects of the human desire to connect, understand, and communicate effectively. The activity was publicized on the city's official advertising channels on Facebook (Appendix P), demonstrating how relevant this type of action is for community engagement. These findings highlight the diverse and nuanced ways in which different audiences perceive and value technological solutions for fact-checking.

The empathy map co-design activity led to the creation of distinct characters by groups, each reflecting unique values. These activities underscored the necessity of designing technologies that foster a sense of belonging and inclusion, tailored to the specific needs and contexts of the target audience. The insights gained from

these exercises will inform the final design recommendations and development of a fact-checking tool.

Chapter Summary

This chapter focused on our research's technical and empirical results on AI-based fact-checking solutions. Initially, we explored various case studies and technologies to understand their application in generative media, followed by an empirical phase involving stakeholders to integrate their values into the proposed solution. Through detailed case studies, interviews, and workshops, we identified key challenges and methodologies in publishing fact-checking results and examined the role of generative AI and chatbots. Additionally, we assessed the practical implications of virtual characters and deep learning in fact-checking. The insights gained from these activities informed our co-design workshops, which evaluated technological probes using the Value Sensitive Design methodology. These findings underscore the importance of trust, transparency, and user engagement in developing effective fact-checking solutions.

Freedom would be not to choose between black and white but to abjure such prescribed choices.

(Adorno and Horkheimer 1997)

CHAPTER 5: Value-frameworks and Discussions

In this chapter of the research, we shall analyze the results obtained in the preceding phases, their implications for proposing a final solution, and the guidelines for the decision-making processes required when adopting the VSD methodology to assess an application designed to counter health misinformation. Initially, we will elucidate how we evaluated the mappings of values, norms, and requirements that were established during the various stages in which our methodology was applied. The diagram below (Figure 61) summarizes the critical milestones in this process, which we will explore in further detail.



Figure 61. Key points of our research within the VSD phases in which they were planned

In the first segment, we will examine the iterative transition from abstract to concrete concepts and back. It is crucial here to recapitulate significant moments from the analysis of cases and the insights derived from stakeholder collaboration during the workshops. The juxtaposition of values and norms will be pivotal in providing a grounded and innovative analysis of the processes outlined. In the second phase of this chapter, we will address the challenges associated with designing a technological solution that is value driven. For this purpose, the values identified will be transcribed into the principles we have highlighted, which are essential even for other potential research implementations in the future.

Finally, we will delineate the limitations and prospects for further work, acknowledging the achievements not realized by the thesis deadline. We will also summarize how the research questions have been addressed throughout this investigative process.

Section 6. Values, Norms and Technical requirements

There is an inherent complexity in the challenge of designing a solution based on values that are often intangible. The key difficulty lies in concretely conveying the aspirations of those defining these values. However, given the complexity of misinformation, these ambiguities have become an intriguing part of the process. This began when we presented stakeholders with the challenge of considering the benefits and drawbacks of using methods similar to deepfake to combat deepfake itself. Finding balance and equilibrium proved to be a way to reach consensus that could better materialize these subjectivities.

During our engagement with a highly multidisciplinary group from various countries, for instance, journalists engaged in debate with experts in technology and regulation at our second workshop, yielding interesting results. The tension between the speed that an AI solution can offer must be weighed against the labor displacement harms that such technologies might cause, and in these cases, the conflicts are not very congruent. As illustrated in the diagram (Figure 62) summarizing these discussions at the workshop, the list of enumerated harms is larger than that of benefits. It is expected that these categories of experts would

exhibit a more conservative bias given their professional and academic choices. However, it is also noteworthy that in discussing benefits, they highlight the engagement that the solution could offer—a significant issue currently faced by communication professionals striving to disseminate relevant scientific content.

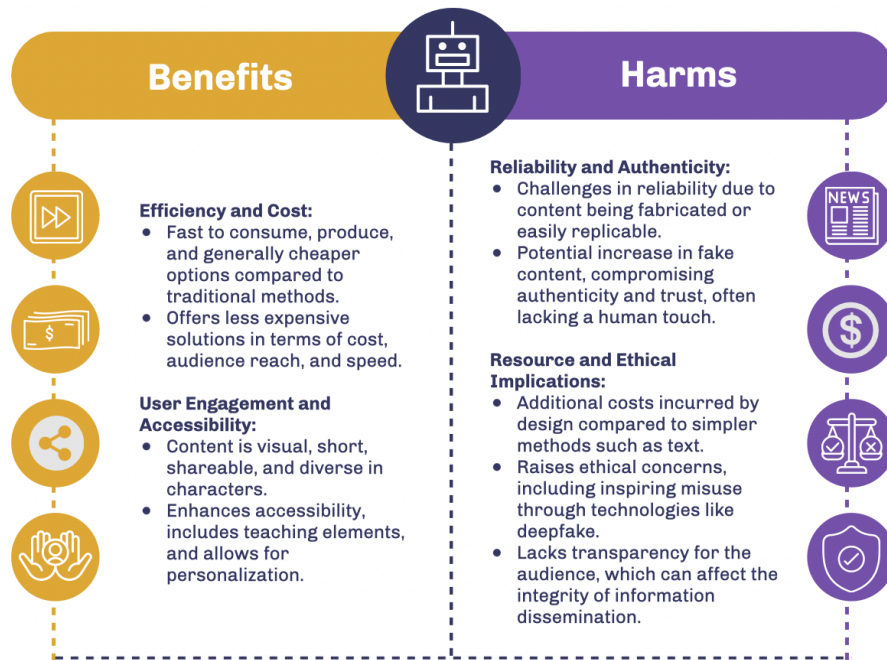


Figure 62. Summary of benefits and harms from the stakeholder presentation at the Zurich workshop

It is also important to emphasize that by highlighting the harms, stakeholders remind us that the lack of transparency in how productions were made could pose another problem if it is not clear to the public that the solution involves manipulation with deepfake. This was proven when we conducted a workshop with elderly caregivers without providing much explanation about what deepfake entails, as there was no similar questioning from them. The impression we got was of a misunderstanding about whether they truly comprehended that reporter Dina Aguiar, for example, had not genuinely recorded that video. Such misunderstanding could be problematic in a real-world solution, as this confusion could further hinder the ultimate goal of communicating validated, truthful information. Therefore, exercising caution and careful consideration of these endorsements is indeed a challenge that requires reflection. Below we will detail how other conclusions about these concepts can guide us towards a proposal.

6.1 Integrating values into the design process

As previously outlined, our mapping was done with a co-participatory empirical approach that incorporated various practical exercises, adhering to the methodological framework of integrating values into technological development. A particularly notable technique we emphasized was using a "values deck." This tool enabled participants to collectively highlight and select values that they deemed most suitable for our specific solution, which envisioned the use of AI to share fact-checking. The "values deck" offered a diverse array of possibilities, allowing for a broad spectrum of perspectives and approaches to be considered. The differences in responses are clearly illustrated in the comparative diagram of our first two workshops (Figure 63), which involved stakeholders from the identification, production and support categories. The first was held in Portugal at the University of Porto and the second was held at Tamedia's event for the Computational Journalism conference in Zurich.

The diagram also helps us to see the values and norms used in the triangulation of data, which were gathered from the analysis of empirical responses in these workshops. Campbell and Fiske (1959) describe that various sources identify its beginnings in psychology and the multi-trait-multi-method matrix as an interest in the triangulation of different quantitative and qualitative data (D.T. Campbell and Fiske 1959). The collection of quantitative and qualitative data is updated when using this approach. According to the literature, the data is used to validate the accuracy of the analysis, as performed in our research.

The first workshop, described in the left-hand column, focused mainly on residents of Porto, Portugal, while the second, described in the right-hand column, included a more diverse group of researchers from various countries. This comparison highlights both the divergences and similarities in the results emanating from these sessions. In terms of similarities, both workshops underscored core human and operational values such as access to information, freedom of expression, democracy, and innovation. These values reflect a shared commitment to individual liberty and social justice. Additionally, there was a consistent emphasis on decision-making norms and emotional skills, including the balancing of public and private interests, making decisions based on verified facts, and

and detailed approach, likely influenced by the diverse backgrounds of the international participants.

It is also interesting to note that some of these concepts do not appear to have been taken into account in some of the solutions observed in case analyses research phase. For instance, the 'Fact Check Assistant' created by Canadian Social Media, despite being intriguing and offering technological development that meets the demand for quickly disseminating fact-checking results, does not seem to adhere to the values of 'informativeness,' 'transparency,' and 'reliability.' This is because its checks were based on a general database consulted by GPT-4, which, in 2023, did not reference sources and had a reported issue with hallucination.

The TITAN case, despite its interesting approach of raising questions, also seems to lack concern for 'engagement' and 'inclusion' in its methodologies. It does not incorporate any strategy to genuinely interest someone in downloading an app to access a fact-checker that challenges their certainties. Nevertheless, this would always be a significant challenge in a complex problem like misinformation but considering aspects such as 'engagement' appears to be a weakness of this solution.

The original solutions from AIDA and GAB are closely aligned with values such as 'equality,' 'innovation,' 'inclusiveness,' and 'respect' (with AIDA inspired by the first black woman to lead the 'Jornal Nacional,' which is an admirable tribute of the solution). However, the criteria for news recommendation were very generic and technical in their initial project, solely focused on textual character count and random ranking, which does not validate aspects such as 'Meaningfulness,' 'Trustworthiness,' and 'Informativeness.' Thus, these reflections on values, although challenging, reveal an intriguing path for how new technologies might be invented or readapted in a society where design increasingly proves necessary to be considered beyond the financial interests of humans alone, for example. In the next section, we'll try to understand how these contrasts between values were viewed and how we can think about this balance.

6.1.1 Identifying and Solving Value Tensions

One method we adopted to examine the tensions among values was the use of participatory workshops with activities that asked stakeholders to express their views on controversial situations. In their first round, the audience primarily consisted of stakeholders from the 'identifying,' 'supporting,' and 'production' categories. The tension was largely focused on traditional journalism being challenged by this innovation. The specter of job displacement is often present in these discussions, a topic that is both relevant and pertinent to the current times. Below, we will highlight how this issue was debated within the groups.

During the final presentation, participants discussed their group's project's central values and norms. They highlighted public health, liberty of expression, right to access information, equality, individual liberty, and democracy. They also identified tensions related to the public interest versus private interest, the push of innovations beyond regulatory boundaries, and the balance between fast response and accuracy in fact-checking. Additionally, they mentioned the potential misuse of technology by the state. In Table 31 are their comments on the final presentation.

Table 31. Asserts from co-design workshop# 1

Sub-themes	Participant	Asserts Example
Ethical data use and decision-making	PT2	The tensions identified regarding this would be the public interest versus the private interest, innovations always pushing regulation boundaries, and fast response precision with innovation.
Speed of response	PT1	Disinformation spreads quickly, so the response must be quick and accurate. Innovation can be problematic as it requires checking and presenting reliable information.
Disclosure of processes and policies	PT3	The last tension would be the state using technology as a deception.
Digital Accessibility	PT3	We can use technology to facilitate fact-checking by using a plugin like those used for translating sign language. This allows us to provide verified information using a 3D model.

During the final presentation, Group 2 discussed the challenges of establishing credibility with AI-generated content in journalism. They emphasized the need for professional curation and highlighted the values of liberty and tolerance. They also explored the idea of creating new possibilities of ethics and morality while using

technology for fact-checking. This new relationality and imagination could call for new forms of ethics to be reflected upon in a vision that is pos-humanistic and beyond the human (Table 9).

Table 32. Asserts from co-design from group 2

Sub-themes	Participant	Asserts Example
Timely delivery of services	PT4	Gaining credibility from robots is difficult, but AI can be a valuable tool with human curation to enhance journalism. Media literacy and innovation are essential to consider."
Ethical data use and decision-making	PT5	We explored the possibilities of new ethics and moralities, particularly in synthetic data with no embodiment, where the main difference between deepfakes and 3D characters lies. It prompts us to consider new ethical dimensions.

During the final presentation, Group 3 emphasized the importance of respect, independence, reliability, and responsibility for their solution. They discussed using technology for fact-checking while ensuring a real human and close relationship with the audience. It is important to emphasize that, despite the problem presented with fact-checking and the slowness of its process compared to fake news, they decided that they would like neither a character nor a deepfake (Table 33). Reliability and security were highlighted as critical factors, especially in the context of deepfake concerns.

Table 33. Asserts from co-design group 3

Sub-themes	Participant	Asserts Example
Reliability and consistency	PT7	While technology can assist with fact-checking, a real person should deliver the information to establish a more human and closer relationship with the audience. Reliability is crucial, considering the issues with deepfake, and information should be delivered independently.

It is intriguing to observe that these analyses are consistent with some uncertainty regarding the moral standard to be adopted in a future where misinformation appears to broaden tolerances. This became apparent in the second round of the workshop, attended by a diverse international audience but primarily comprised of fact-checkers, normative professionals, data analysts, and AI developers from the "identifying," "supporting," and "distributing" categories. The

submission from the third group, led by PW9, presented tensions concerning another value they identified: informativeness. It intersects with central norms like providing explanations and citing sources. These values often conflict, as we discussed various scenarios and trade-offs. Table 34 present theirs and others identified values and norms.

Table 34. Asserts from co-design workshop# 2

Sub-themes	Participant	Asserts Example
Ethical data use and decision-making	PW6	The primary tension we discovered exists between privacy and transparency. Balancing the need to protect an individual's identity, such as that of a journalist, while also ensuring content attribution to maintain transparency, poses a challenge. We also discussed the use of watermarks, which users should recognize when AI has been involved in content creation. Implementing watermarking or metadata for AI-generated content is crucial for identification.
Clarity and comprehensibility	PW7	Our primary values encompass several vital elements, with privacy being paramount. With the use of deepfakes, there's a risk of using people's images without authorization, infringing upon their privacy. Inclusiveness is another core value, as we observed a bias towards perfection. We need to assess if our pursuit of perfection aligns with inclusivity. Additionally, we prioritize ease of use to ensure user-friendliness. To achieve this, designers and engineers are working on various aspects. Transparency is crucial as well, especially in anti-fake news applications. We need to ensure transparent processes for logical reasoning and enhanced fact-checking. For instance, attributing monkeypox to COVID-19 in a video without proper explanation raises concerns. Therefore, we must focus on delivering transparency in a technically sound manner,"
Accuracy of health-related information	PW9	Transparency can be challenging for the audience at times. We also explored the distinction between deepfakes and 3D models, as deepfakes involve real people imitating something else, whereas 3D Avatars are common in computer games. Task 2 revolves around the importance of inclusion as a core value, along with norms related to accessibility and representation. We need different versions of content, including videos, audio, and platforms, to cater to diverse audiences on platforms like TikTok, Instagram, and Twitter. These efforts contribute to our overall commitment to inclusivity.

Somehow, we perceive a resistance to deepfake technology due to its negative associations with the very name by which the technology has become known, linked to misappropriation. However, as previously explained, the technology itself is not necessarily used solely for this purpose. Its use could be legitimately authorized, for example, allowing a figure like Dina Aguiar to have a contract permitting her image to be used to disseminate AI content from data that come from fact-checking agencies she has authorized. The idea of using experts to validate the

use of deepfake technology seems rather abstract to journalists who are accustomed to fighting against their images being discarded in a highly volatile market.

I do not dismiss this difficulty; however, from my point of view, there are other severe issues in this replacement of labor, such as the following: what would be a fair salary to pay to a Dina Aguiar for having her image associated 24 hours a day, seven days a week with an automated service? If there is a new problem of surplus value created by AI, it is urgent that these professionals organize themselves to understand the problem from within, with all its nuances, rather than merely adopting a reactive and prohibitive stance. Especially when we realize that, from the point of view of our last workshop, aimed at elderly caregivers who represent more of the "distributing" category of our stakeholders, Dina Aguiar's deepfake technology was the best ranked overall by the participants. This way of choosing an avatar that validates an established and legitimized figure is an interesting result of our research. Innovation was balanced against other values by this group of participants, as interviewed highlights below in Table 35.

Table 35. Asserts from co-design workshop# 3

Sub-themes	Participant	Asserts Example
Representation of diverse groups	PQ3	I believe that those who do not have internet access or do not know how to use the internet are excluded from this solution. From the ages of 60 and 70 onwards, many people do not know how to operate the internet.
Availability and accessibility	PQ10	I would like to teach caregivers how to access this. That's my opinion, because most informal caregivers do not have access.
Creativity and originality	PQ8	For five years, I transported ill people. I had two experiences with some Alzheimer's patients. One, a familiar person. I used music in the car every day. I even recorded music. We would play guessing the singer. They would say the name. Look, singing is not my forte. But they sang along. The music greatly improved the Alzheimer's patients, the man was transformed.
Digital Accessibility	PQ9	Sometimes we don't have access to a doctor. And I would like to have information to show to people. To have it right there. To have access like a person. Yes. To have it very clearly explained, showing the sources, it could be useful.

From the perspective of stakeholders directly impacted by misinformation, the tensions are somewhat different. Regarding "equality" as a value, the technology

would not be beneficial if they do not know how to use it, but learning about its usage and teaching it is appealing to them. They also perceive the use of sound and image differently, in a more constructive manner, influenced by the loneliness that can accompany the later years of life. Here, artificiality seems to be seen not only as a component of intelligence but also as a form of artificial sensitivity that our inhumanity may have somewhat neglected. Given such rich insights, in order to connect with our empirical experiences, the next session will elucidate key principles and discussions concerning the values, norms, and requirements previously identified. The list is not exhaustive, but it is a contribution to a thorough research effort seeking some connection to proposing a path.

The thematic analysis of data extracted from participatory workshops revealed a broad range of considerations categorized into three main groups: Neutral, Positive, and Negative. These discussions led to the identification of eight key codes, which were selected as fundamental principles for the proposed solution. The neutral discussions primarily focused on the informational aspects and the importance of accuracy and rationality in fact-checking technologies, particularly within the context of public health. Participants emphasized the necessity of ensuring that information is easily accessible to the public, promoting a logical and evidence-based approach, considering the public health implications of disseminated information, and ensuring the completeness and comprehensibility of the information. These considerations underscored the importance of reliability, trustworthiness, and transparency to ensure that fact-checking technologies are effective and dependable.

Positive discussions highlighted the potential benefits of automated fact-checking technologies, emphasizing the need for intuitive and easy solutions to protect user privacy, engage users meaningfully, and ensure that the content is relevant and valuable. These positive aspects highlighted the values of efficiency, innovation, engagement, and meaningfulness, demonstrating that technological solutions should not only be efficient but also capable of engaging and meeting users' needs in a significant way.

Negative discussions addressed the ethical concerns and potential risks associated with the use of deepfake technologies and fact-checking, including ensuring that the technology is inclusive and respectful of all demographic groups,

maintaining freedom of expression and ensuring content authenticity, considering aesthetic norms and how they may affect public perception, and protecting democratic values and ensuring equality and justice in information dissemination. These concerns underscored the importance of inclusion, respect, authenticity, and freedom, emphasizing the need for technological solutions to respect and promote these fundamental principles.

The analysis revealed a clear need to balance efficiency and innovation with ethical considerations and fundamental values. By integrating these eight key codes, the proposed solution aims to be robust, ethical, and effective, meeting both user needs and expectations for reliability and transparency. These principles will guide the development and implementation of fact-checking technologies based on deepfake, ensuring they can combat misinformation ethically and efficiently.

6.1.2 Mapping the values

The thematic analysis of data extracted from participatory workshops revealed a broad range of considerations categorized into three main groups: Neutral, Positive, and Negative. These discussions led to the identification of eight key codes, which were selected as fundamental principles for the proposed solution. The neutral discussions primarily focused on the informational aspects and the importance of accuracy and rationality in fact-checking technologies, particularly within the context of public health. Participants emphasized the necessity of ensuring that information is easily accessible to the public, promoting a logical and evidence-based approach, considering the public health implications of disseminated information, and ensuring the completeness and comprehensibility of the information. These considerations underscored the importance of reliability, trustworthiness, and transparency to ensure that fact-checking technologies are effective and dependable.

Positive discussions highlighted the potential benefits of automated fact-checking technologies, emphasizing the need for intuitive and easy solutions to protect user privacy, engage users meaningfully, and ensure that the content is relevant and valuable. These positive aspects highlighted the values of efficiency, innovation, engagement, and meaningfulness, demonstrating that technological

solutions should not only be efficient but also capable of engaging and meeting users' needs in a significant way.

Negative discussions addressed the ethical concerns and potential risks associated with the use of deepfake technologies and fact-checking, including ensuring that the technology is inclusive and respectful of all demographic groups, maintaining freedom of expression and ensuring content authenticity, considering aesthetic norms and how they may affect public perception, and protecting democratic values and ensuring equality and justice in information dissemination. These concerns underscored the importance of inclusion, respect, authenticity, and freedom, emphasizing the need for technological solutions to respect and promote these fundamental principles.

The thematic analysis of the workshops revealed a clear need to balance efficiency and innovation with ethical considerations and fundamental values. By integrating key codes, the proposed recommendation aims to be robust, ethical, and effective, meeting both user needs and expectations for reliability and transparency. These principles will guide the development and implementation of fact-checking technologies based on deepfake, ensuring they can combat misinformation ethically and efficiently.

Summarizing so many concepts and rich discussions is no easy task, but we take Floridi and Cowls' initiative as an example, when they started a debate on condensing the ethical principles of AI into five key concepts, four of which were derived from classic biomedical principles: beneficence, non-maleficence, autonomy and justice. In the end, they introduced a fifth principle, explainability, as a new concept to facilitate AI. As part of our contribution, we decided to document these conclusions in a table with 8 values, standards and requirements that represent the main considerations and principles that the groups identified in their debates. They cover ethical, practical, and technical aspects of designing virtual characters and content in the context of deepfakes and related technologies and serve as a framework for future projects and initiatives:

Table 36. Author's Ethical Matrix into Fact-Checking Disclosure Research Insights

Value	Norm	Requirement
Trust	Attribution of AI-generated content	Implementation of watermarking or metadata for content identification
Engagement	Captivating Content	Design and engineering efforts to enhance engagement factors
Information	Clear Explanations and Cited Sources	Detailed explanations and proper source citations to ensure informativeness
Inclusion	Accessibility	Multiple content representativeness versions for different accessibility needs (videos, audio, platforms)
Novelty	User-Friendliness	Continuous improvements in design and usability
Promptness	Information Accuracy	Fact-checking mechanisms and clear sourcing of information
Aesthetics	Conformity to Societal Standards	Ensuring avatars content does not inadvertently reinforce societal beauty standards
Equality	Protection of Personal Data	Robust encryption and data security measures

The framework we will propose in the next section encompasses principles within these points and focuses on ethics, stakeholders, and technology. Regarding Ethics, we will address the theoretical foundations of the meta-design phase needed to incorporate values into the design process and methodology and take responsibility for the resulting products or services. In Stakeholders, we focus on the research and synthesis phases of co-design workshops. The third pillar, Technology, covers the ideation and evaluation phases of the design process. This segment provides our proposal tool with embedded values in mind. The general objectives are to analyse and critically evaluate the impact of our research project.

6.2 Ethical Guidelines and Principles

In addition to assessing the value of technological evidence presented during participatory workshops, the value framework developed in this dissertation provides insights into designing future health fact-checking applications. This section describes various principles and recommendations to support the application of AI in the fight against misinformation. Initially, a set of value-based guidelines generated through empirical research is presented. Subsequently, technological design recommendations are illustrated through a mock solution. The

principles selected followed the thematic grouping of the workshops data already presented and are Trust, Engagement, Information, Inclusion, Novelty, Promptness, Aesthetic, and Equality.

6.2.1 Principle of Trust

Here, trust is considered fundamental for users to rely on the capabilities of a health fact-checker. Regarding health, the phenomenon of trust requires special attention because, unlike disinformation, sincere doubts can lead both to belief in fake news and to changes in opinion about an initial verdict. Therefore, focusing on trust is crucial in our proposal. This principle emphasises the accuracy, reliability, and consistency of the checker in delivering true health news. Building trust involves implementing rigorous verification processes and ensuring that the checker's sources are credible. For this reason, our recommendation strongly relies on a database that directs to various established fact-checking channels, such as the COVID-19 Fact Checkers Dataset developed by Ryerson University for the Social Media Lab (2020).

The COVID-19 Fact Checkers Dataset, curated by Ryerson University, encompasses a compilation of over 200 active fact-checking entities focused on verifying misinformation concerning COVID-19, developed in collaboration with the World Health Organization (WHO) as part of a global initiative. The choice of data curation is the main reason we recommend a system that uses a fact-checking database rather than the model of some LLMs like ChatGPT, as done by the Canadian Fact Check Assistant project, described in our literature review.

Although companies like OpenAI are working on creating link references in their content generations, and users are beginning to have access to real-time news reports ("OpenAI" 2024), the lack of transparency in the criteria for this and the issues already described about hallucinations in these models makes it not recommendable for a solution like the one we are proposing. "Trust" was also the reason why we adopted the use only of LLM models for the creation of textual and visual language and not in the search for knowledge curations. We believe that changing this paradigm - from strengthening fact-checking to not valuing the power of technology in searching for random misinformation using LLMs - is an important

and significant difference in our research compared to other emerging technology product launches.

This is also an interesting point achieved with the use of the VSD methodology, which ensures that the solution is thought from values, and not the opposite, of already having a technology ready and from that using the social sciences for its ethical validation. This is a recommendation we point out in our research, that at this challenging social moment we live in, we may change this paradigm and rethink whether some AI solutions should be suggested from a prior assessment of our values and not the other way around.

6.2.2 Principle of Engagement

Engagement refers to the ability of our checker to maintain user interaction. Our recommendation is for an engaging solution that not only responds efficiently but also presents information in an interactive and appealing manner, encouraging continuous use and interaction. This analysis point was central in many of our co-design activities and survey questions. We understand that the item of familiarity was important in this context, to help us reflect that the checker should operate in a way that is intuitive for users.

It should use language and interfaces that are easy to understand and navigate, reducing the learning curve and enhancing user comfort. Our research has shown that when making a solution suggestion, the workshop participants in Macedo placed characteristics of a known person, that is, working with reference people can be an interesting use. However, the possibility of customization brings other interesting points, which is the phenomenon of resonance. Here we recall the work of Rosa (2019) highlighted in our literature review, that the fast pace of contemporary life leads to problems exacerbated by the spread of fake news, but that a possible solution lies in achieving "resonance". This possibility of customizing the avatar or voice according to the user's own aspirations is a highlight of what this technology can offer and should be highlighted in the fight against misinformation, as it allows for some insertion in the message, an interesting factor in terms of confirmation bias in the dissemination of content by digital media.

Resonance, as he describes, is a form of open and reciprocal relationship with the world, where the individual feels that the world speaks to them in a personal and meaningful way. However, the constant pressure to be more efficient and the relentless race against time make human relationships more instrumental and superficial. As a result, people feel disconnected, even though they are more "connected" than ever through technology and media. Creating ways for people to connect more with content can then be a way to improve resonance in the face of alienating speed.

Another interesting point that we highlight as a design recommendation for these technologies are the negative factors such as frustration and fear that should be avoided. Thus, a very aggressive approach like that of the grumpy cat should be avoided and a character like that of the friendly cat was better received because it minimised frustration and avoids feelings of irritation or dissatisfaction. The checker should provide quick and accurate responses and have capabilities to handle and gracefully recover from errors or misunderstandings. This involves careful consideration of whether the chatbot interactions generate even more fear about the subjects, ensuring that it does not overwhelm or mislead users about health issues, which can be sensitive and potentially alarming.

6.2.3 Principle of Information

The principle of information underlines the role of the checker in providing comprehensive, accurate, and relevant data. Here we reiterate the recommendation for our solution to be used to speed up data checked in fact-checking organizations. This is a recommendation aimed at using AI to reinforce the importance of journalism and not to replace it, a recurring criticism that journalists print on AI projects. It is worth reinforcing that at this moment we are prioritizing journalism and the fundamental values described in our workshop as fundamental precisely because stakeholders representing fact-checking agencies, journalists, regulators, and university professors from the media were present.

In the context of a health misinformation verifier, access to information refers to the ability to obtain correct, relevant, and high-quality health information. Information should be readily available to the public and understandable for people

with different levels of health literacy. The source of the information should be clear and transparent, such as government agencies, health organizations, or scientific publications. We recall that in our research the recommendation is that our verifier should be equipped to offer detailed explanations and additional resources that enhance the user's understanding and decision-making. Regarding this, we highlight fundamental journalistic values that are rationality, informativeness, and source grounding. Informativeness is essential to provide clear and comprehensible explanations about why a claim is true or false.

For this, it is important that the explanation is clear, so that anyone can understand it; without omitting important information; and contextualized, providing information that allows the public to understand the subject in question. The citation of sources refers to the evidence supporting the verification of a claim. In the context of a health misinformation verifier, this includes scientific evidence, concrete data, such as statistics and verifiable facts; and the citation of recognized authorities, such as government agencies, universities, and health organizations. Therefore, it is necessary in the implementation that fact-checking agencies bring these elements in their verifications.

6.2.4 Principle of Inclusion

Inclusion involves designing the verifier to be accessible to a diverse range of users, accommodating various languages, cultures, and abilities. This ensures that all users, regardless of their background, can benefit from the service. Thus, technology can be beneficial for its automatic availability in different languages if it starts from this criterion. Of course, the recommendation for inclusion starts from some limitations on how these technologies are built, and thus the difficulty of including different dialects in automation and the cost of training an algorithm based on certain languages that are not relevant for commercial purposes. And precisely for this reason, linguistic inclusion is another fundamental aspect to be included as a recommendation in similar projects.

Tools of this type, with support for multiple languages, allow individuals who do not have the native language as their first language to better understand health verifications. When we recommend inclusion we know that some scenarios will be

more difficult and perhaps costly, but perhaps it is precisely for these scenarios that we should highlight projects that include these technologies in certain cultures and niches where we need to reduce the AI divide. This is exactly what we tried to do by going to an economically disadvantaged region of Portugal in part of our tests and the results in Macedo showed the demand they had for inclusion when freely proposing a technology they chose the representation in the voice of a radio journalist, which is possibly the most impactful way of reaching news in that place.

We also reinforce the recommendation that these technologies be resonant to the place, open to meanings and senses that they may have in each scenario of action. This also refers to a reflection still on the segmented and customized use that can be made of 3D characters. Technology, then, should reflect that this customization of the verifier user has the freedom of inclusive choices in skin colour, genders and age. The use of text-to-speech (TTS) technology has the potential to promote inclusion in accessibility for people with disabilities, as it can be a crucial tool for people with visual impairments, dyslexia or other reading difficulties, allowing these groups to access digital texts more efficiently and promoting digital and academic inclusion (Smith & Jones, 2020).

In the case of blind or visually impaired people, the tool allows them to hear texts that would otherwise be inaccessible. For those who have dyslexia, hearing texts aloud can help them better understand the content. Similarly, voice recognition technology (speech-to-text) can expand access to knowledge and opportunities, facilitating communication and reducing barriers. This is a recommendation for use to precisely meet the demand in our Macedo workshop, when suggesting a character in the empathy map exercise, users suggested feeling listened to as a radio reporter does. This inclusion in digital platforms can be promoted by integrating text-to-speech tools in this application to facilitate the receipt of the user's question from the fact checker, making these environments more accessible.

6.2.5 Principle of Novelty

Incorporating novelty in the verifier involves integrating innovative features that distinguish it from other chatbots, but beyond that, that draw attention to the

content and language used in the dissemination of checked news. This includes advanced natural language processing capabilities as a factor to be highlighted in our proposal in terms of technology to be applied by our solution. Generative media, in terms of audio generation, demonstrated good acceptance by users and did not bring great resistance in balancing with other values. Within this, Natural Language Processing (NLP) models such as BERT, GPT (from OpenAI), and other models that can understand and generate human language efficiently, are recommended for this solution.

We highlight here the good results with prompts that generated the creative use of grumpy or kind characters, an interesting capacity of the innovation that can be applied with the use of these technologies. In terms of generative media, we highlight the use of computer vision, with computational techniques in image generation, specifically focusing on Neural Radiance Fields (NeRFs) and Transformer/diffusion models, such as Dall-E. The generation of 3D characters was also an accepted suggestion in our research, when our participants iterated from our use of a cat to the use of an owl as a character for communication. We also highlight the capacity of this technology to enable means of generating an interesting scenario automatically. In our research, our prompt requested that the API generate the scenario of Trás-os-Montes, with characteristic mountains and vineyards, and the result was quite similar to real scenarios, which represents an interesting innovation to be applied in a service that aims to encourage user engagement. The characteristic of deep learning to be used in image replication, known as deepfake, also demonstrated good receptiveness by the public, both in the option where it was employed with the use of a well-known reporter, and the one that was used with an actress in the synthesis tool.

However, due to other strong impacts of the tension with values such as equity, which we will detail further below, we do not recommend this type of use for a solution like the one applied to fact-checking. It is worth noting that this is also because we consider there to be other technologies as described earlier in this item, or even creation using photogrammetry, that can have a good impact and help achieve the expected result.

Another point to highlight technologically is the possibility of creating customizations and giving the user the freedom to make these "assemblies" of

characters and their constructions in terms of the gender of the voices to be selected and phenotypic characteristics. Still as a way to create resonance and inclusion, we recommend that this parameterization may have the possibility to use or not use image, voice or text, in order to include blind, deaf people or those who for more personal reasons do not want to use all the possible multimedia resources. The research showed there to be divergences in the final format, but the great advantage of this technology is precisely in allowing the diversity of choices and our function is to potentiate engagement and not to impose some kind of arbitrary definition of ours as to the form of the message.

6.2.6 Principle of Promptness

We describe as an important recommendation that the verifier have the capability to deliver information quickly. In the fast-paced world of health news searches, timely updates and quick responses are essential to maintain relevance and usefulness. This capability is a fundamental technological differentiator that offline means cannot provide in combating disinformation, and thus we understand it as a capability that justifies and balances the use of these AI technologies proposed for a solution to be suggested. Firstly, the immediate availability of responses allows users to receive instant feedback, crucial in situations where incorrect health information can quickly spread and cause panic or distress.

This immediate access to verified and accurate information significantly improves the user's experience, who seeks reliable answers in moments of uncertainty. Here it is worth highlighting the inserted factor of "sincere doubts" in the spread of misinformation about health. Moreover, the capability of automatic response of these systems allows the service to scale to handle a large volume of interactions, essential in periods of public health crisis when rumors and false information can proliferate rapidly, and also a problem that we saw to be great in the pandemic infodemic of Covid, for example. By offering consistent and precise responses in an automated manner, these systems can reach a much larger number of people simultaneously than would be possible with manual checks.

The 24/7 nature of these chatbots is another valuable characteristic, providing continuous support, regardless of the time or geographical location of the users. This is particularly important for users in different time zones or in situations where access to health professionals may be limited. The constant support increases the public's trust in official information sources and can help mitigate the spread of misinformation in critical moments.

Finally, the generation and analysis of data from user interactions with the verifier offer valuable opportunities for continuous improvements. These data can be used to refine the system's algorithms, ensuring that automatic responses remain relevant and effective in the face of changes in misinformation trends. Thus, in addition to improving the quality of the service, data analysis helps adapt the verifier to proactively respond to emerging health fake news, ensuring that the public receives reliable and scientifically based information.

This is a benefit that the EducAGE project could have in its content portal, for example, because if some misinformation was not found they would have access to longer form content to be developed to deal with the issue more consistently. It is worth highlighting here that alienation, as presented by Hartmut Rosa, has a factor intimately linked to a certain speed of society in which it implies that people do not connect meaningfully with the message, as they understand that everything is ephemeral and meaningless. Hartmut Rosa explores the idea that the social and technological acceleration of modern life contributes to a profound alienation of people in relation to the world around them. According to Rosa, this speed prevents the relationships of resonance that are essential for a meaningful life experience. Rosa argues that this alienation not only impoverishes the individual experience but also weakens the social fabric, as authentic and deep connections are replaced by ephemeral and performance-oriented interactions. Therefore, using promptness to combat the superficiality of speed can also be a necessary way in a society like ours.

6.2.7 Principle of Aesthetics

Here we highlight the relationship with the visual and functional design of the verifier interface. An aesthetically pleasing design can increase user satisfaction and

contribute to the overall usability and acceptance of the message itself. In this sense, our recommendation is on the diversification of options and the use of technological resources to provide a more pleasant experience and some personal resonance with the technology. One recommendation is the use of 3D characters, which can be essential for creating visually engaging and accessible educational materials. These characters, when used in photogrammetry scenarios, allow for the creation of eye-catching visuals that reduce the more negative content of explanations about complex procedures or health conditions.

This not only improves the public's understanding of health issues but also makes medical education more inclusive, effectively reaching people with different learning styles. Furthermore, as we experienced in our interactions, the use of these 3D characters and personalized awareness campaigns can be highly effective in communicating public health issues. There was no great repulsion to the use of animals in our approaches, for example, which justifies us suggesting that these solutions can be culturally adapted and translated into various languages, ensuring that the messages are relevant and impactful in different community and cultural contexts. This is particularly useful for addressing specific health problems that may be prevalent in certain regions or among specific population groups.

On the other hand, we also saw the potential use of DALL-E, with its ability to create visualizations based on textual descriptions, which can help in the creation of fact-checking materials that are culturally adapted and linguistically accessible. This facilitates the understanding of complex health information, especially for those in communities where access to quality health education is limited. We highlight here a very interesting observation made in our research that care must be taken not to reinforce aesthetic standards and that there must be diversity in the approach to characters, characterizations, and ages to be used. We understand that this diversity is not easy to obtain, but treating communication in more playful terms can be interesting.

6.2.8 Principle of Equality

The principle of equality ensures that the chatbot provides the same level of service and information to all users without discrimination. This involves unbiased

algorithms and equitable access to information, reinforcing the chatbot's role as an impartial tool for news verification. Within this principle, our decision was to recommend that the verification be based on content pre-selected from fact-checking agencies, which although they also have limitations and errors, are based on investigative journalism and values of informativeness and really combating misinformation.

Furthermore, is precisely our reinforcement that this type of solution extends beyond content creation and into areas like education and misinformation management. The verifier serves not just as a technological tool but as an educational ally, aiding citizens in discerning accurate information, crucial in the healthcare domain. Another interesting insight about equity in our project was observed when caregivers offered their help to teach the elderly how to use new technologies, such as tools not only to help correct misinformation but also for equity in digital literacy.

Another interesting point of recommendation is in the observation that these solutions can serve to reinforce the need for AI tools that are culturally sensitive and ethically aligned. In journalism, particularly in health education, respecting and understanding cultural nuances is as important as providing accurate information. Still within this, the EducAGE project Verifier proposed demonstrates how AI can be effectively utilized in regions that might not traditionally be at the forefront of technological innovation.

By focusing on specific needs and leveraging AI for educational purposes, such initiatives can help bridge the global AI divide, offering models that can be replicated in other socio-economic contexts. We conclude the recommendation on equity by reinforcing transparency in the solution we want to propose and this is due to being open about how the chatbot works, particularly how it obtains and verifies information. This clarity helps build trust and credibility, as users understand the mechanisms behind the news verification process.

Therefore, there must be clear mechanisms of accountability to deal with errors or failures of the chatbot, and users must clearly understand that they are interacting with a chatbot and not a human. This becomes very relevant when we observe in our research, for example, that it is not clear to our participants that they are dealing

with a manipulated image of Dina Aguiar, and not the real Dina Aguiar. Therefore, the process by which the verifier reaches its conclusions must be transparent.

Another suggestion also recommended in the workshops was to create a watermark in our health fact-checking system. First, we recommend using this watermark in all communications, including text only, or audio and videos. This would allow users of the verification system to quickly recognize and confirm the authenticity of the content. In addition, it would be beneficial to develop a tool within the verification system capable of detecting these watermarks, ensuring the authenticity of the content in real-time, which is especially useful for content that circulates widely on social media and other digital platforms.

6.3 Mock-up proposal

Considering these challenges, we recommend how generative media could be used in a balanced way to publish and share fact-checking content, more precisely, in the challenging misinformation ecosystem. Continuing the discussion on technology applied to fact-checking, we present workflow contextualized within value-oriented empirical principles (Figure 64).

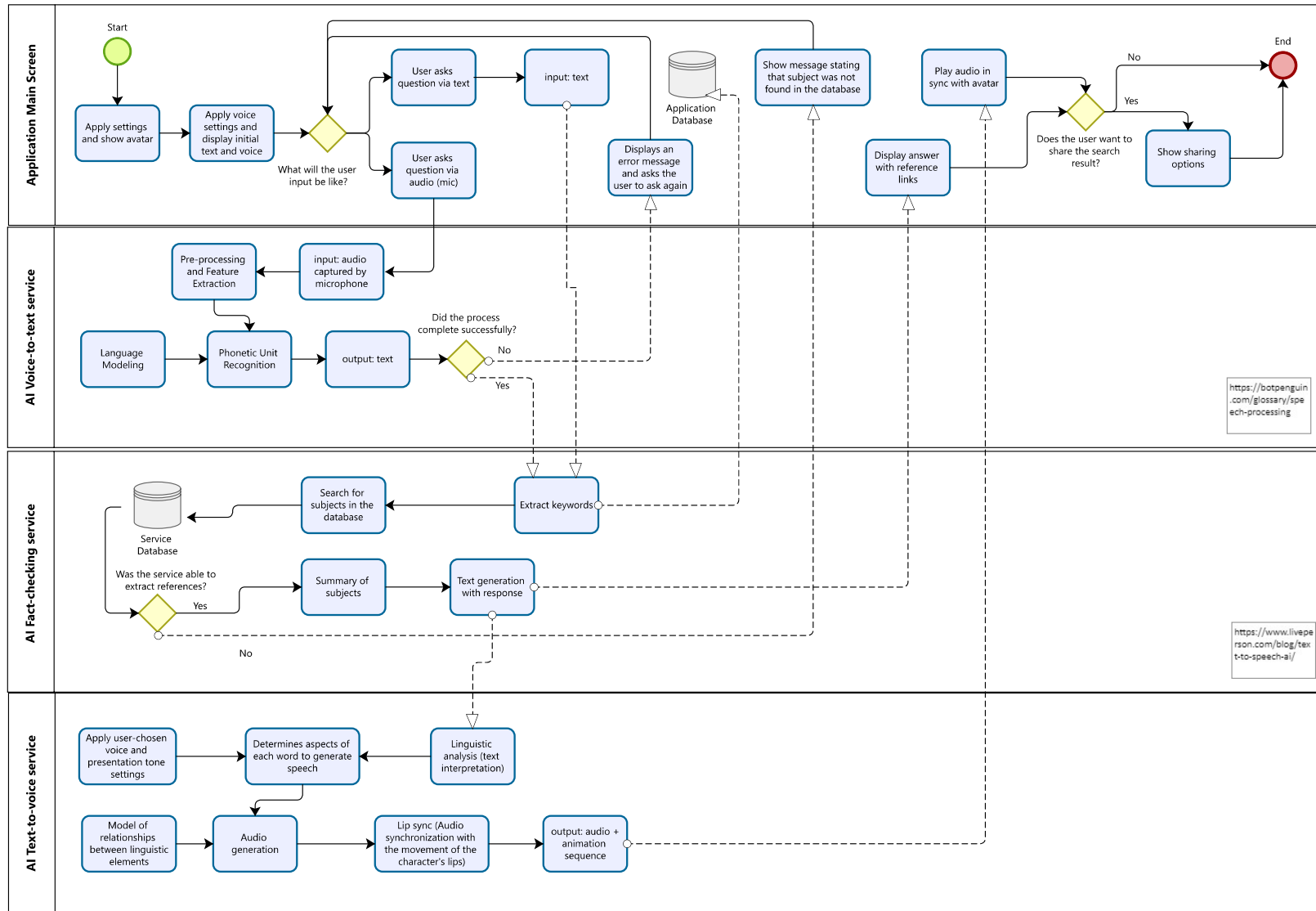


Figure 64. Verifier main screen workflow created by a collaboration between Haline Maia and Priscila Ariel (Globo)

We recommend an intuitive user experience, with action elements and the tool's main features easily accessible, as presented by the configuration screen flow (Figure 65).

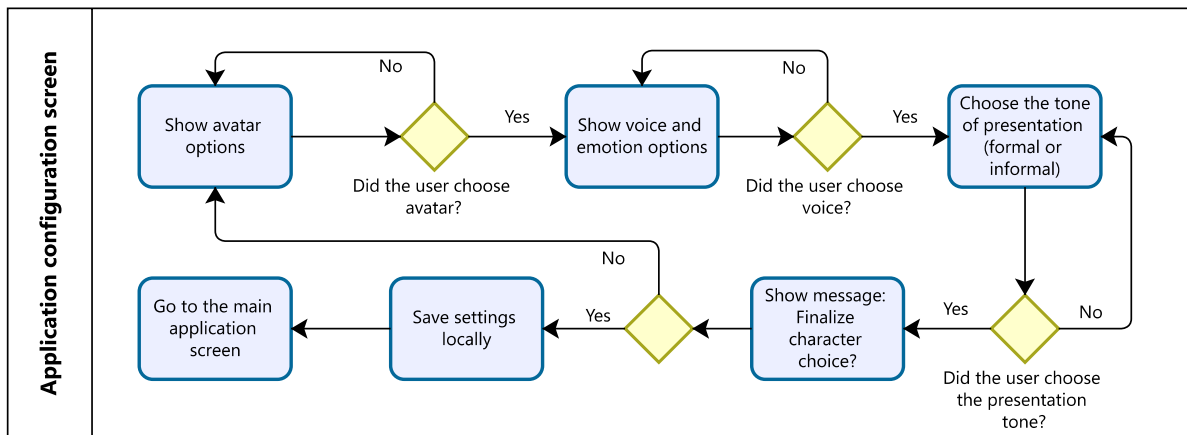


Figure 65. Configuration workflow allowing users to customize the verifier before accessing the main application screen

The synthetic media generation within fact-checked content follows seven steps (Figure 66). Users can initially choose remote access to the interface, with customization options and potential sharing features. The automatic fact-checking system then selects the most relevant content, which undergoes a summarization process to condense the original, credible news. In step 3, the summarized text is converted into audio using AI. Subsequently, lip synchronization (lip-syncing) is performed. In step 5, the scene is rendered, and the character presents the news according to the user's preferences. Finally, in step 6, a video is generated in real time.

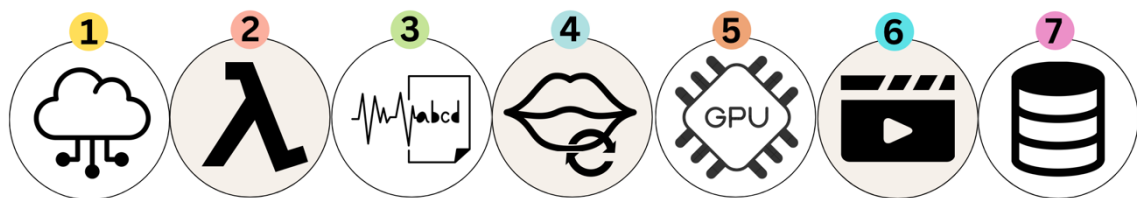


Figure 66. System workflow for our verifier synthetic generation proposal

Our proposed virtual verifier addresses the governance challenges in combating health misinformation. Recognizing this complexity, we aim not to solve the problem entirely but to offer a simple yet useful tool. Our proposal is a plugin based on our experimentation in Macedo de Cavaleiros. Still, it is designed

to be agnostic and applicable to various portals and scenarios, leveraging the potential for technology to co-evolve with social structures. The following outlines the basic initial steps for such a solution:

- Step 1: Remote access to the system interface allows users to type their query or use speech-to-text technology via mobile phone or laptop. This aligns with our research findings on "Inclusion," highlighting the importance of users feeling "heard." Users can also customize their avatar, choosing between a pet or human, selecting the voice without predefined gender based on visual elements, and setting the tone (e.g., formal, informal, serious, kind, or grumpy).
- Step 2: An automated fact-checking system selects relevant original fact-checking texts. Here, we rely on a predefined database, such as an existing fact-checking repository. In a summarization technique called LSA - Latent Semantic Analysis (Bottari and Ariel 2020).
- Step 3: AI transforms the short text generated in the previous step into audio. Once texts are created, a Text-To-Speech (TTS) service can be used in the voice module. It implements the so-called "neural" voices based on deep learning and results in a realistic-sounding voice.
- Step 4: Lip synchronization animates an avatar, with options including a human, a pet, or just text and voice.
- Step 5: The scene is rendered, and the avatar presents the news according to user preferences. Steps 4 and 5 shall occur in real-time. The reference system uses a particular instance equipped with a Graphics Processing Unit (GPU) and a local webserver to orchestrate the process.
- Step 6: The video is generated or streamed in real-time, with options for sharing on social media.
- Step 7: A report detailing the verification process and the algorithm's response is generated. This report helps evaluate system efficiency and identify unanswered queries, enabling the development of long-form content for the hosting portal. This approach does not aim to solve all misinformation problems but contributes to mitigating them.

The Verifier System allows users to search for fact-checked health misinformation and obtain summaries of verified hoaxes. The tool aims to help users verify rumors

and fake news, providing and sharing information. The software's core value lies in offering a friendly, dynamic interface, contributing to the current technologies combating misinformation. In step 1, users access the interface, designed for usability across various devices (mobile phones, tablets, computers). A live mock-up from these workflows (Figure 67) for the recommended application applied to the already existent Educage App can be reached by entering the link:

<https://cloud.justinmind.com/usernote/prototype/c5befb1cdc6d94b144f91664c27a19dd4addf58f70a4065425859f0283a00ac2>

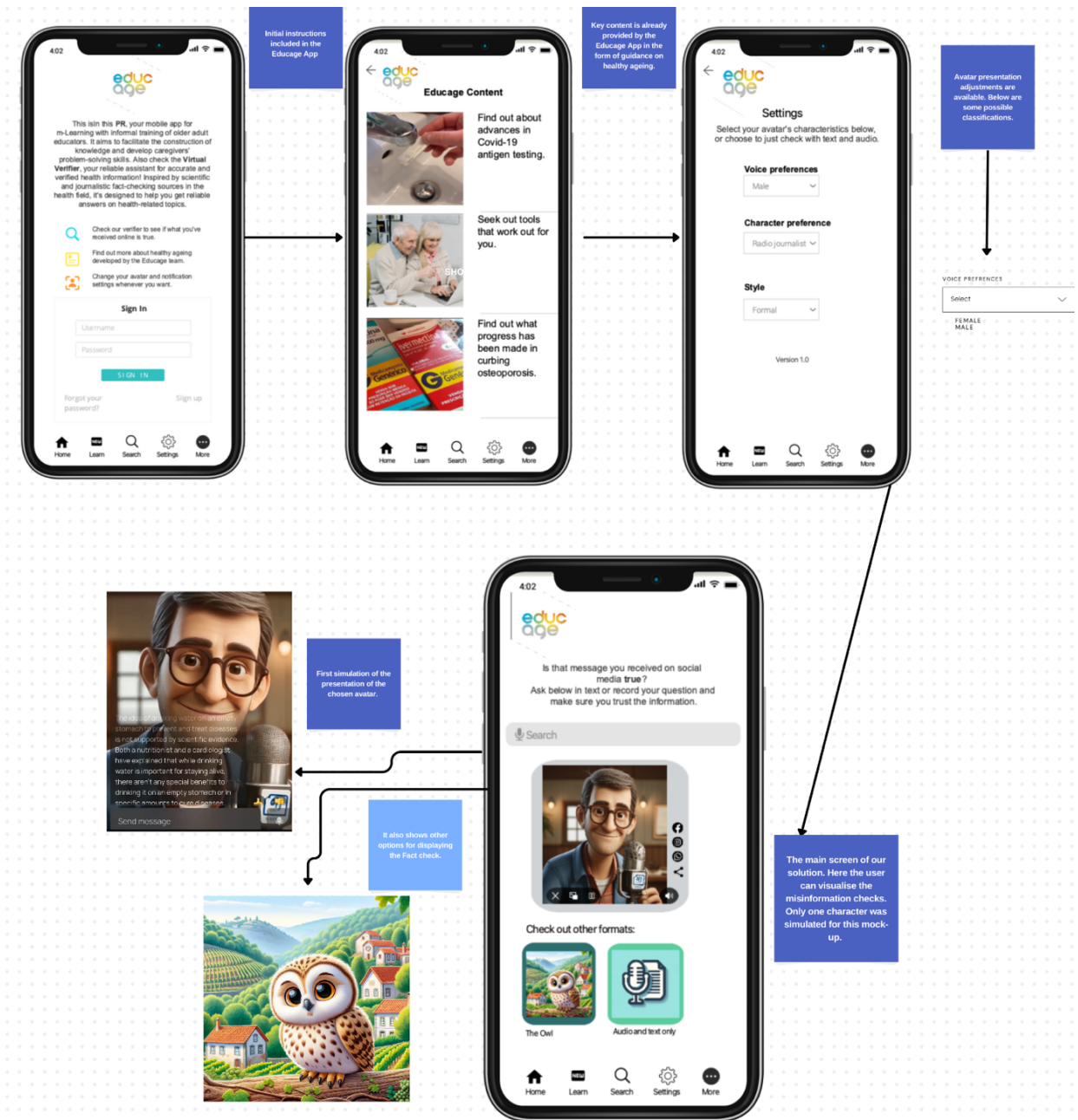


Figure 67. Screen Flow about how our users will experience the platform on mobile

We have developed the same version for the Moodle site Educate. It's important to note that our plugins are designed to be inserted into other channels that provide health information, such as Educage or others that could be a channel for accessing this information or even fact-checking organizations. We reinforce this is an important point in our proposal since we believe that an application like this alone would not be downloaded, nor do we see governance

for this type of solution. Inclusion in other website, which already have credibility and trust, could help the proposal to be useful and thus resolve the difficulties in the rapid dissemination of health misinformation verifications, an answer to our main research question. This is the live mock-up (Appendix Q) for the web version:

<https://cloud.justinmind.com/usernote/prototype/b9b4ca314648516623fcc1dc9b07ec144b0cdb2b01dc177d6088073dfe8ceb97>

An important highlight of our empirical research was the inclusion of a watermark in the AI-generated video, as well as a gadget on the site from which it is generated so that people who have received this content via their networks can check that this video was indeed generated by our verifier. This is an interesting solution for generating trust and credibility for the project, and we have put it forward as a strong recommendation. Still, here in this mock-up it is not represented, as we believe it would be something for further research in the future and more focused on implementation. Another interesting point for implementation would be to think about how exactly the system's message would be constructed, as here we only propose in general terms of "formal," "affectionate," or "grumpy." The appendices G show the prompts we generated in the LLMs used for the experimentation phase. Still, it would be interesting to do more research into the form of the message used in implementing the solution, even with a more ethnographic approach with cultural analysis. The time and funding limitations meant that we had to stop here but with good prospects for further research. We'll explain more about this in the limitations section.

6.4 Limitations of the Proposed Framework

The research presented in this manuscript offers a comprehensive approach to developing value-informed design recommendations for a fact-checking verifier, yet it acknowledges certain inherent limitations that are crucial to recognize. These

limitations illuminate the complexities and challenges associated with applying a value-informed design framework in practical settings. The study also touches upon a broad theme which has been shown to interact with diverse audiences. Although the challenges were known at the outset of the research, the seriousness of the issues at hand, coupled with personal reasons previously outlined, provided the impetus to continue focusing on this area. Therefore, it becomes necessary to discuss the limitations encountered further.

Primarily, the value framework detailed in this dissertation is structured to address infodemic scenarios where misinformation can profoundly affect public understanding and behavior. However, it is within these scenarios that a greater number of fact-checking agencies are found producing data, which forms the basis for our solution. Nevertheless, this focus may restrict its applicability in contexts where different value tensions and stakeholder priorities might arise. While the framework excels in data-rich and user-interactive scenarios, it may not adequately capture the nuances required in less dynamic environments.

Another significant limitation concerns technology governance. The framework presupposes a certain level of governance and oversight, which may not be universally present across all organizational or geographical contexts. This variance can pose challenges to the practical implementation of the recommendations, as different regions and sectors may have varying regulations and compliance requirements that could hinder uniform application. Moreover, addressing a problem like misinformation, which is both ubiquitous and nebulous, becomes challenging. For this reason, our proposed solution involved seeking additional partners to assist with the ongoing maintenance of a solution like ours. This approach was directly linked to this limitation; thus, the continuation of a project like ours is also dependent on sustained partnerships with other civil society initiatives.

The dissertation also recognizes the difficulty of fully representing underrepresented groups. Despite efforts to integrate diverse perspectives, the depth and breadth of these representations may still be insufficient. This is critical as it could impact the framework's ability to serve all potential users adequately, particularly those with unique or severe disabilities who might require more

specialized considerations. We also face limitations related to time, the challenges of multidisciplinary, and budget constraints.

The co-evolution of technology and social structure is another noted constraint. The framework advocates iterative design and adaptation; however, aligning technological advancements with evolving social policies and structures can be slow and meet with resistance. This misalignment can obstruct the effective implementation of using a deepfake of a known journalist like Dina Aguiar, as we could not establish a fair wage for her image to be available around the clock. We note the need for further research in this field and the implementation of discussions on the additional value of wages to avatars under these circumstances.

Thinking long-term and at scale presents another hurdle. While the framework aims to address immediate design needs, scaling solutions to a broader audience and maintaining their relevance and effectiveness over time can be challenging. Long-term sustainability requires continuous funding, stakeholder engagement, and updates to the technology, which may not always be feasible.

Lastly, concerns about AI surveillance in journalism, particularly about user data privacy and security, pose ethical and operational challenges. As the framework involves collecting and analyzing user data to enhance widget functionality and user experience, ensuring data protection and ethical use is paramount. This issue is especially relevant given the increasing scrutiny over data privacy in digital solutions. Nonetheless, we encourage journalists to continue studying and understanding AI so that the discussion is not superficial and merely prohibitive, as we know that disinformation-generating initiatives are not so cautious.

In conclusion, while the research provides a solid foundation for designing a fact-checking solution sensitive to citizens' values and needs, these limitations underscore the need for ongoing evaluation, stakeholder engagement, and adaptive strategies to ensure that the framework remains effective and relevant in various contexts. We understand that there are implementation challenges, but we have gathered consistent data to map out and guide the way forward in an ethical and viable manner.

Chapter Summary

This chapter examines the application of Value Sensitive Design (VSD) methodology in evaluating a technological solution to counter health misinformation. It analyses results from preceding phases, highlighting the mapping of values, norms, and requirements established during the study. The chapter begins by elucidating the iterative process of translating abstract concepts into concrete applications, emphasizing significant insights from stakeholder workshops. Stakeholder engagement revealed tensions between the speed of AI solutions and potential negative impacts, such as job displacement.

Methods for resolving value tensions through participatory workshops are explored, highlighting principles such as trust, engagement, information, inclusion, novelty, promptness, aesthetics, and equality. A framework of value-based guidelines and technological design recommendations is proposed, underscoring the importance of ethical, practical, and technical considerations. The chapter concludes by acknowledging the limitations of the proposed framework, including the challenges of fully representing underrepresented groups and aligning technological advancements with social policies.

Nonada.
O diabo não há!
É o que eu digo, se for...
Existe é homem humano.
Travessia.⁴

Grande Sertão: Veredas,
Guimarães Rosa (1908 - 1967)

CHAPTER 6: Conclusions

In this investigation, we analyzed the use of artificial intelligence technologies to combat misinformation in the health sector. We concluded that we met our initial research goals by adopting a method centered on seeking technological applications whose critical analysis was related to the ethical challenges posed by the applications of generative media and the recent challenges of misinformation production. In this sense, we managed to answer our research questions with a final proposal that was only suggested after deep reflection.

The search for a final tool proposal during these times of pandemic trauma could not come solely from systematic reviews without the particularity of seeking citizens' opinions; therefore, our trajectory had a significant participatory component. Engaging in dialogue with citizen science is a challenge shared by modern society, trapped in a complex network of causes and effects generated by health agents, journalists, technology developers, and citizens in general. As a result, this thesis had an exploratory nature, with an approach capable of giving voice to different stakeholders and seeking collaborations within this complex

⁴ The closing passage from Guimarães Rosa's book *Grande Sertão: Veredas*, a literary work that explores the complexities of human nature and the Brazilian hinterlands. The following is a loose translation: Nonada. There is no devil! That's what I say, if it's true... What exists is human humanity. Journey.

ecosystem. In the following pages, we present the main testimonies of this research, which directly reflect on each research question and initially presented objectives, as well as other assertions that emerged during this process.

The vast scope of this research presented significant challenges, prompting us to narrow the topic to address the critical importance of health misinformation fact-checking, especially highlighted by the onset of the COVID-19 pandemic. This dual focus aimed to create a meaningful social impact and leave a lasting legacy through our research. To achieve this, the methodology was designed to focus on the problem at hand while maintaining a diverse and inclusive dialogue with various stakeholders. This approach ensured that the research remained grounded in real-world applications and perspectives, thereby enriching its relevance and potential for impactful solutions in combating health misinformation using advanced AI technologies.

We began our journey by reflecting on some classical authors of communication theory, focusing on the digital revolution in journalism and exploring its ontological foundations and paradoxes within the media and technology ecosystem, especially in the context of fact-checking and post-truth. We also examined how technological innovation has always driven significant changes in journalistic communication, from the transition from manuscripts to print facilitated by Gutenberg's movable type to the influence of radio and television. It is interesting to observe from empirical research with observations reported by journalists in our co-design workshops how these theories still make sense and how these media challenges are cyclical.

In the literature review chapter, we also saw how the epistemological path of cyberculture shifted from an optimistic view to a more pessimistic one and how incorporating digital technologies transforms communication and the dissemination of information. Initially, with Pierre Lévy, we highlighted "collective intelligence" as an emerging cyberspace phenomenon where knowledge is collaboratively constructed and distributed through digital networks. The technological realism advocated by Barbrook, in contrast, criticizes the economic and exclusivist controls over technology, highlighting how power uses futuristic theories to obscure real problems. Within these dichotomies, it is important to

emphasize that since its inception, the Internet has had its critics, and the very popularity of the term post-truth directs us to understand the difficulty in establishing truth and denotes this epistemological trajectory.

Regarding disseminating health misinformation, especially during the COVID-19 pandemic, we highlighted in our literature review the conceptualization of the phenomenon of misinformation and the importance of the emergence of fact-checking. Likewise, we understand that the divergence over the validity of fact-checking is a symptom of theoretical novelty, making disputes over its pertinence or not remain. In academic debates, another important point we could highlight is that there are different versions of the role of journalism in pre-bunking done by fact-checking depending on the cultural context and the ideological spectrum of the theorist, with the difficulty of neutrality in science being no novelty for the humanities, which we once again observed in our research.

After a broader evaluation of cyberculture and the definitions of misinformation and fact-checking, the second chapter advanced to a more specific understanding of how the digital ecosystem is characterized by a complex interaction of media technologies, including AI and machine learning, which have the potential to improve fact-checking and content generation. At this stage of our research, it was important to understand AI literacy to address possible collaboration misunderstandings for the final proposal of technological solutions. In this part, we included the conceptualization of authors such as Allaham and Diakopoulos, who were also important in helping us understand how AI models like GPT are used to generate text, although they still face challenges such as "hallucinations" or the production of unintentional false information.

At this point in our literature review, we observed how the practicality and instrumentality of technology are evident in how society organizes knowledge and how AI has a significant social component when applied in journalism. Simondon's work helped us understand the complexities of technology, tools, and machines, emphasizing their purpose and contribution. With a clear and detailed understanding of AI technologies, we aimed to make a better and more appropriate contribution to the field of human-computer interaction (HCI) when proposing an approach to attributing value to the use of generative media during health

misinformation, precisely because the VSD methodology is anchored in this major area. Nevertheless, Cohn & Cope highlight understanding how the economic context directly influences access to information technology, with more affluent populations having greater access, and this dynamic underline the disparities in technological reach and utilization.

This connection led us to a relevant part of our research, where we addressed the ethical challenges of using Artificial Intelligence. Here, we had an important reflection on the social challenges, moral dilemmas, and the misuse of AI with deepfakes. It is worth noting how the use of deep learning was driven by a negative connotation in the very semantics of the word "deepfake," which automatically makes us associate it with its misuse to deceive and manipulate people for individual interests. At this stage, we also highlighted the reflection on the dichotomies using Memes and the interesting conceptualization of Hartmut Rosa on resonance and alienation in current times. Reflecting on these concepts was extremely relevant to the conclusion of the thesis.

Another significant moment of the research was the incorporation of inoculation theory, which posits that exposing individuals to weakened forms of an argument can strengthen their cognitive defenses against subsequent attacks. Roozenbeek and Van Der Linden, by advocating the preventive exposure of individuals to deceptive techniques, fallacies, or misleading information, provided us with an interesting theory to address in our methodology so that our stakeholders could evaluate in our workshops the possibility of using deepfake to combat deepfake. This was a turning point in finding better approaches to generating fact-checking that aimed to create an informed and resilient public capable of critically evaluating information and protecting themselves against the influence of misinformation. This conceptualization was important for the subsequent empirical phase.

In ethical considerations, it is worth highlighting our search for a model that could guide us in designing an ethically responsible technological tool, which was one of our research hypotheses. In this sense, we highlight the choice of the Value Sensitive Design (VSD) methodology, whose concepts were relevant to ensure that human values are considered in the technological design process. At this point, we

highlighted the identification of value tensions as an important theory and anchored the idea defended in the foundations of this methodology of how data ownership provides power.

Given these ethical frameworks, the adopted perspective explored the relevance and importance of fact-checking to provide verified and valuable data in infodemic scenarios. The theoretical systematic review in this thesis identified a limited number of studies on the automation of fact-checking content publication, with only 9 percent of the studies in our review specifically researching this topic between 2020 and 2022. Our research presented at the International AAAI Conference on Web and social media at the TrueHealth 2023 Workshop, highlighted the relevance of this finding to the academic community. In the same literature review, we identified tools like ClaimReview, developed in partnership with Google, promoting verified information by displaying fact-checked statements and validity notes, thus combating the rapid dissemination of misinformation. We also observed that the technology for automatic fact-checking is a promising area with the application of models based on BERT, convolutional neural networks (CNNs), and supervised machine learning, offering a new perspective for this field. However, there are still a few tools that could use generative media for fact-checking dissemination.

In this sense, our methodology incorporated the three VSD phases to be explored: conceptual, theoretical, and empirical investigations. These phases served as a parameter and impetus for the methods that were adopted throughout the data collection process. Additionally, we contributed by adapting the methodology to divide the stakeholders to be consulted into direct and indirect categories as defined by UNESCO in its disinformation study. This helped us create a rationale for narrowing our scope in such a broad and diverse ecosystem. The idea of addressing the categories of identifiers, supporters, production, and distribution—terms already defined by the UNESCO study—served as a parameter for saturation in our qualitative research.

Chapter 4 seeks to answer our research hypotheses through a detailed thematic analysis of the central themes addressed in the thesis. The entire content of the interviews, workshops and focus groups has been transcribed and is attached in

full as an appendix to the thesis. These transcripts were manually coded into the proposed themes in order to summarise them in the conclusion, using NVivo 14 software to quantify the coded themes and create a graphical summary (Figure 68). This treemap chart allows us to visually represent the proportions of key themes from the qualitative data, which are misinformation in Health, Artificial Intelligence, Fact checking and AI Ethics, as shown in the treemap chart. These proportions reflect the emphasis placed on each theme, based on the qualitative data collected.

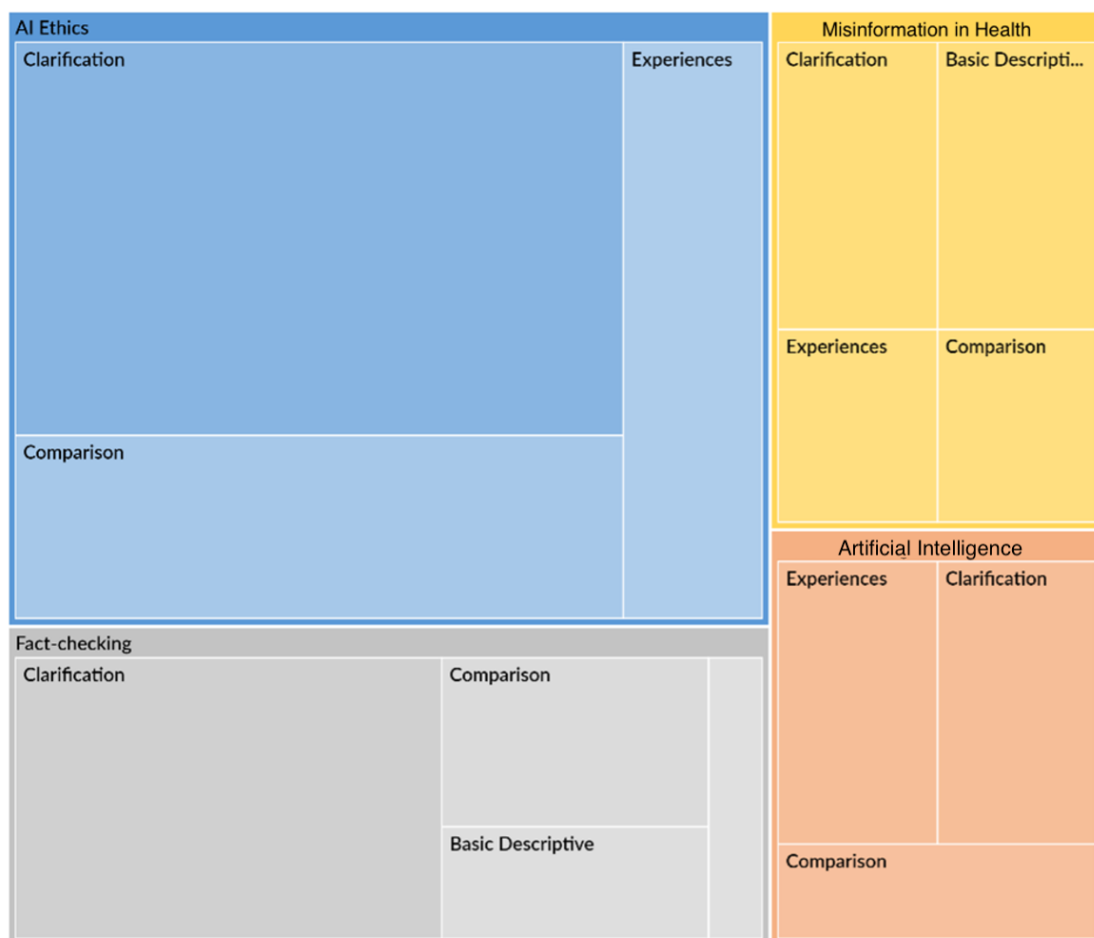


Figure 68. Distribution of Responses in AI Ethics, Fact-checking, Artificial Intelligence, and Health Misinformation

Starting with misinformation in health, represents a significant portion of the thematic analysis, highlighting concerns about the spread of misinformation and the role of AI in addressing these issues. Artificial Intelligence accounts for a smaller yet important share of the thematic focus, indicating significant interest but somewhat less intense than AI Ethics and Fact-checking. Responses here focus on

explaining the favorable aspects of automated solutions, outlining the differences and practical impacts.

AI ethics, representing most of the thematic focus, indicates a deep engagement with the ethical considerations of using AI in health fact-checking. Discussions likely covered the need for ethical considerations, the tensions and benefits of these ethical implementations, and real-world insights or incidents highlighting ethics' importance in AI applications. This suggests a strong emphasis on understanding and defining ethical principles and practices.

Moving on to fact-checking, which also takes up a substantial part of the responses, it suggests that the effectiveness and challenges of AI-enhanced fact-checking tools were a key focus. This includes detailed discussions about the preferred solutions, the essential elements of these solutions, comparative assessments of their value, and user experiences that shed light on practical impacts.

Within the thesis totality, the sub-themes include Experiences, Clarification, and Comparison, with Experiences being the largest sub-theme. This suggests a strong emphasis on real-world applications and user experiences with automation. Overall, the treemap chart reveal the depth and distribution of discussions across these critical themes, highlighting the areas of greatest concern and interest within the context of AI application in health fact-checking.

This research aimed to broaden the investigation into the responsible use of generative AI within journalism, more specifically, fact-checking in health. It was designed to address the lack of information in this field, especially understanding how various stakeholders receive and assimilate content developed using AI to help in acute health misinformation scenarios. This was followed by research into the perspectives of 30 individuals interviewed within the pre-selected categories of identifiers, supporters, production, and distribution in some phases of individual interviews, focus groups, and workshops. Each of these categories was: identifiers with 5 participants, distribution with 14 participants, supporters with 8 participants and production with 5 participants. Based on these conclusions, a proposal of fundamental principles and a mock-up of a technological solution were developed.

These interviews, literature reviews and case analyses provided the data to answer our hypotheses and research objectives.

The **first research question** (How can integrating artificial intelligence in disseminating fact-checking outcomes contribute to mitigating misinformation related to health on digital platforms), coded “Misinformation in Health,” was answered in various ways. We observed cases like the “Fact Check Assistant” created by Canadian Social Media, detailed in our case analysis session, developed especially for health issues at a Hackathon at Toronto Metropolitan University using Chat-GPT, a similar solution focused on helping to contain the chaos caused by misinformation and presented in the same workshop where our research was presented titled Combating Health Misinformation for Social Well-being-Disinformation.

We also spoke with elderly caregivers in Macedo de Cavaleiros, served by the Educage project in a European consortium created to develop digital tools to aid in health development. A noteworthy response to this contact was when our audience reinforced their interest in our solutions and demonstrated that they were unsure whether some of the content verified by our prototype was true or false, finding it a useful solution. They even detailed the previous use of media tools for integration with elderly patients in treatments such as Alzheimer's. Nonetheless, they primarily showed enthusiasm for the proposals and the exercise of choosing their characters' attributes, which was a highlight of our research. This participatory co-design activity, in which they articulated their ideas on paper, simulated the functionality of an interactive technological solution in which users could personalize their avatars. This process was helpful for proposing a technological tool that allows personal choices to resonate in an AI-powered tool.

A particular point of the research reinforces the answer to this research question with a phrase said by a participant in response to an inquiry about the relevance of these technologies when they have simple questions and real misinformation doubts: "Because the doctors, you know? They don't inform the patient; they don't say what should be." It seems evident that we as a society still have much to do in building bridges with valid information for citizens. Neglecting the use of technologies in infodemic scenarios is a luxury that we communicators do not have

if there are many information deserts worldwide, making the use of technology necessary, useful, and much more applicable in the scope of artificial affectivity than our titular artificial intelligence proposes about the machines we create.

The **second research question** (how can artificial intelligence facilitate disseminating authenticated health information) identified with the code “Automation” was also answered through our case analysis, complemented by an in-depth interview with Globo's technology and innovation researcher Priscila Ariel, developer of the AIDA and GaB prototypes and specialist in the subject. Understanding the details of this solution and the entire context of its insertion was fundamental for a proposal that rethought the use of these technologies in a viable and relevant way. Still, in collaboration with the researcher, we could effectively and compellingly simulate these solutions by creating tests presented to the various stakeholders who participated in our research.

This experimental process was presented through a well-structured and comprehensive approach. The collaboration with the developer of Globo's AIDA and GAB prototypes and using two generative media APIs ensured the exemplification of how these technologies are viable and can reach various scenarios. Still, we highlight that using these models made it more possible to evaluate user feedback and the effectiveness of synthetic media technologies in generating content for disseminating fact-checking. Moreover, adapting the prototypes for disseminating digital fact-checking demonstrates the versatility and adaptability of the proposed solution and its ability to be customized in various scenarios.

It is also worth noting concern about the existing APIs on the market since our project did not intend to end with an implementation; we decided to experiment with some that were available. In this context, we could experience some limitations in the available tools. It should be noted that these APIs have limitations in terms of gender, age, racial characteristics, etc., even though some try to demonstrate some representativity. For example, when recreating Dina Aguiar through the free Akool tool, it was impossible to present a result close to the journalist regarding age and body composition. The actress with the closest characteristics used in the deepfake modeling made Dina Aguiar younger,

reinforcing what our collaborative workshop participants warned about using aesthetic standards and reinforcing stereotypes that avatars could offer. We recognize that this may only be a fact due to using the API for rapid prototyping in the free version and that paid versions may have more options, but we must highlight this potential problem in these software tools to be corrected by future developers.

The **third research question** (in what ways can the process of health fact-checking benefit from the incorporation of artificial intelligence to speed up the dissemination of verified information on digital platforms) represented by the code "fact-checking" was also extensively scrutinized. Here, we highlight again the literature review that was important to provide a conceptual foundation of the functional basis, criticisms, and relevance of this action in the face of health misinformation challenges. This analysis was deepened with case studies like the Full Fact agency, which has initiated campaigns to analyze COVID-19, and the TITAN project's detailed approach employing the Socratic dialogue approach. Our first empirical approach, especially noteworthy, brought a public of journalists and fact-checking specialists who, inserted in this context and challenge, expressed their approval for technologies to be used for various stages of the misinformation fact-checking process, recognizing the usefulness of use in dissemination once results have been validated by credible bodies. As one participant highlighted, observing automatic translation in various languages: "We can use technology to facilitate fact-checking by using a plugin like those used for translating sign language."

Here we also highlight the fact that researchers investigate less about the promising field of generative media for disseminating fact-checking, with 91% of the research observed in our systematic PRISMA review conducted for the monitoring, spotting, and checking phases. With this validated, we highlight that our research put into practice the use of automation in the publication phase, conducting various tests with technological proofs using deepfake technology through the Synthesia and D-ID APIs, in addition to simulations performed with pre-existing Globo prototypes that were not previously used for fact-checking. In total, we simulated 12 different technological probes for fact-checking

dissemination, some in the more traditional "false" or "true" models used by fact-checking agencies, but others differentiated by using a more affective or fun language like the cats made for the Educage project. The experimentation done using photos, 3D characters, and deep learning adds an interesting dimension to the research, allowing the exploration of different approaches to content creation. This enriches stakeholders' experience and facilitates meaningful discussions about the benefits and challenges of using synthetic media technologies.

Overall, our research presented an innovative investigation into the potential of synthetic media technologies for disseminating fact-checking. The use of technological proofs and comprehensive experimentation demonstrates a high level of scientific rigor, making the study highly valuable for advancing the field of synthetic media and its applications in content generation and fact-checking dissemination. We believe that the consistency and iteration of this process conducted in three application phases with validation from heterogeneous publics can be considered our most significant research contribution to the presented theme and problem, given the methodological consistency and addressing an identified research gap. Nevertheless, we reiterate the resonance power of these technologies when applied, observing the public's interest in elderly caregivers in performing the empathy exercise of creating their own characters, as well as the lack of rejection of new technologies, with 83% of participants preferring deepfake solutions such as pets and text and audio only, while only 17% still preferred fact-checking dissemination done by a journalist from the Polígrafo fact-checking agency in Portugal.

Finally, the **fourth research question** (how can integrating artificial intelligence promote ethically responsible practices in health fact-checking explicitly focusing on disseminating counter-information) coded as AI-Ethics brought the largest part of our responses. Here, it is worth emphasizing our initial interest that we sought interventions for responsible innovation; therefore, including indirect stakeholders in our approach was directly related to understanding how AI impacts the labor structure of the journalism profession, already fragile in our current society for various other factors. We understand that choosing a final proposal that does not endorse the use of a deepfake of journalist Dina Aguiar, for example, was a

contribution in this sense, given the need for more discussion on this topic of regulating deepfakes in journalism.

We believe that, based on this research, we can point to regulatory and discussion gaps that need to be remedied, and here we reiterate, as we have done at other points in our research, that there is a discussion about the surplus value of remuneration for the data provided for image and voice use projects with generative AI that needs to be better debated. Moreover, the solution made with human-like deepfake, for example, was preferred by 25% of our interviewees. Nevertheless, our final proposal for this research aimed to emphasize that it is possible to proceed with innovation on other terms with shorter steps while we discuss more and try to take control of a society that is good for current generations but also for past ones.

It is also possible to highlight this difficulty when we placed stakeholders from different countries to debate these challenges at the C+J – DataJ 2023 Conference. The surveillance over AI was found in several journalists remarks, such as this one: "We prefer alternatives like audio animations or text over AI-generated content (avatar)." Here, it is difficult to understand strictly if this opinion is based on more specific resistance to the solutions we presented, but in any case, it was quite clear the resistance and discomfort of these professionals consulted with our provocation is being replaced by avatars to replace real people. Other interesting highlights from this co-design workshop were that they pointed out the dangers of these solutions as "hard to rely on because it's fabricated" and that they do not have a human touch. However, making all these reservations, it is clear that the balance of values and a final proposal is very relevant for the moment we live in since the use of AI for generating disinformation with biased motivation does not have the same concern.

We conclude that the Value-Sensitive Design (VSD) approach provided a valuable framework for examining the challenges of disseminating fact-checking results. Our research emphasizes the importance of understanding governance structures, as well as the complex interactions between direct and indirect stakeholders in shaping strategies and directions. Currently, this mission is still in its early stages, lacking comprehensive long-term activities such as outcome

evaluations. In the Portuguese case, our research highlights how grassroots, bottom-up approaches influence practical implementation. AI technologies that enable tailored interactions offer an avenue for citizen participation and engagement on a small scale, but allowing for highly personalized and active collaborative efforts.

It is important to evaluate empirical proposals cautiously, so the final part of our thesis with the value-oriented mock-up proposal was dedicated to a detailed summary of the main points we could extract from a proposal. We also sought to make a workflow recommendation for possible applications, recognizing the difficulty of making proposals with effective and feasible value balance. I would like to follow the recommendation of Batya Friedman, founder of the VSD methodology, that sometimes we should even decide not to use AI after all the value balances or have the readiness to consider this possibility. We believe that this proposal, along with the entire process of iterations during the conceptual and empirical phases, can be considered a finding of our research. With this, we understand we have built a consistent foundation for the final proposal.

The construction of these ethical guardrails serves as a way to discuss the long-term impacts of value-driven technological solutions on public trust in media and technology. This challenge is a scientific impact proposal aiming to envision future scenarios that bring evaluation and reflection values in technological design. It is important to evaluate these conclusions with caution. Looking at present-day phenomena requires humility, as noted in historical studies, as temporal distance improves understanding of facts and illuminates debate. It is also necessary to recognize the inherent fragilities of empirical research. It is a snapshot of the present but is hardly capable of keeping up with globally relevant narratives that require theoretical structures that shape concepts and interpretation models. Nevertheless, we tried to build a research approach that, besides incorporating the voices of different stakeholders, had some diverse and generational cultural representation.

Our empirical phase did not create concepts but served as a thermometer to measure those investigated in our conceptual phase, measuring the real. Thus, the conceptual phase created the rationale to measure what was presented in the field. The broader concept of ethics, being the definition of truth an important subset, is

a construction of language with unavoidable inferences from the customs and cultures of each interviewed group. We end our journey hoping that the contributions presented here will be reassessed with other scenarios and cultural contexts, with a basis that can help them be adapted and implemented.

Therefore, it is a complex task to conclude a project built upon a network of debates and collaborations, knowing that the problem of health misinformation is far from being solved by a single solution. We end with numerous new questions and ideas for the future, yet also with the challenge of synthesizing many learnings into words. I take reference from one of Brazil's most renowned writers, João Guimarães Rosa, whose work comprises approximately eight thousand words formulated as neologisms, a habit arising from his study of over twenty languages, including archaic Portuguese and popular expressions used in the Brazilian hinterlands. His masterpiece, *Grande Sertão: Veredas*, for instance, begins and ends with the term *nonada*, a neologism blend of "non" and "nothing" to denote something of no importance, reflecting eternity.

Recommending innovative approaches of communicating in times of crisis required our seeking for new ways that blend the best of our artificial or natural intelligence but keep our sensitivity centered on what is essential in being human. Writing in times of infodemic madness, as well as by translating the hardships of the Brazilian hinterland, seem to refer to words that don't exist. We cannot precisely determine the most appropriate language to use in the interplay of human language, computer language, and the new era of misinformation. My neologism to summarize this research would revolve around words invented to express how the past and future shape the present, balancing our best values and aiming for a better society amidst the chaos of health infodemics. Now, then, and forever.

REFERENCES

- Abernathy, Penelope Muse. 2023. "News Deserts: A Research Agenda for Addressing Disparities in the United States." *Media and Communication*. 11 (3): 290-292.
- Acemoglu, Daron, and Pascual Restrepo. 2018. "Artificial intelligence, automation, and work." In *The economics of artificial intelligence: An agenda*, 197-236. University of Chicago Press.
- Adamopoulou, Eleni, and Lefteris Moussiades. 2020. "Chatbots: History, technology, and applications." *Machine Learning with Applications 2*: 100006.
- Adorno, Theodor W, and Max Horkheimer. 1997. *Dialectic of enlightenment*. Vol. 15. Stanford: Verso.
- Ahmad, Ahsan, Aftab Tariq, Hafiz Khawar Hussain, and Ahmad Yousaf Gill. 2023. "Equity and artificial intelligence in surgical care: A comprehensive review of current challenges and promising solutions." *BULLET: Jurnal Multidisiplin Ilmu 2* (2): 443-455.
- Akaslan, Dursun, and Mustafa Ulukavak. 2022. "Using Photogrammetry for Modelling Realistic Characters in Virtual Reality." The proceedings of the 4th International Conference on Virtual Reality.
- Akst, Daniel. 2013. "Automation anxiety." *The Wilson Quarterly (1976-)* 37 (3).
- Alhindi, Tariq, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. "AraStance: A Multi-Country and Multi-Domain Dataset of Arabic Stance Detection for Fact Checking." *NLP4IF 2021.*: 57.
- Alkaissi, Hussam, and Samy I McFarlane. 2023. "Artificial hallucinations in ChatGPT: implications in scientific writing." *Cureus* 15 (2).
- Allaham, Mowafak, and Nicholas Diakopoulos. 2024. "Supporting Anticipatory Governance using LLMs: Evaluating and Aligning Large Language Models with the News Media to Anticipate the Negative Impacts of AI." *arXiv preprint arXiv:2401.18028*.
- Almeida, Ronaldo de. 2019. "Bolsonaro presidente: conservadorismo, evangelismo e a crise brasileira." *Novos estudos CEBRAP*. 38: 185-213.
- Almutairi, Zaynab, and Hebah Elgibreen. 2022. "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions." *Algorithms* 15 (5): 155.
- Aloshban, Nujud. 2020. "Act: Automatic fake news classification through self-attention." Proceedings of the 12th ACM Conference on Web Science.
- Amaral, Inês, and Sofia José Santos. 2019. "Algoritmos e redes sociais: a propagação de fake news na era da pós-verdade." In *As fake news ea nova ordem (des) informativa na era da pós-verdade*, edited by Imprensa da Universidade de Coimbra, 63-85. Coimbra.
- Amazeen, Michelle A. 2013. "A Critical Assessment of Fact-checking in 2012." *New America Foundation*: 1-40.
- Amrute, Sareeta. 2020. "Bored techies being casually racist: race as algorithm." *Science, Technology, & Human Values* 45 (5): 903-933.
- Arendt, Hannah. 1967. "Truth and politics." *Truth: Engagements across philosophical traditions* 295.
- . 2013. *The human condition*. Chicago: University of Chicago press.

- Ariel, Priscila. 2023. Personal interview on 5 May 2023. In *Google Meet*, edited by Haline Maia.
- Arnold, Phoebe. 2020. *The challenges of online fact checking*. Technical report, Full Fact.
- Arntz, Melanie, Terry Gregory, and Ulrich Zierahn. 2016. "The risk of automation for jobs in OECD countries: A comparative analysis." *OECD Social, Employment and Migration Working Papers*. 189.
- Arroyave, Jesus. 2012. "Health, news, and media information." *The handbook of global health communication*: 194-214.
- Asaro, P, and W Wendell. 2017. *Machine Ethics and Robot Ethics. The Library of Essays on the Ethics of Emerging Technologies Book Series*. New York: Routledge.
- Ashimova, AB, ES Sultanbayeva, Guo Steve, and AT Belgarayeva. 2022. "SWOT analysis of current state and prospects for the development of artificial intelligence in Kazakhstani media industry." *Journalism Series* 66 (4): 51-58.
- Atanasova, Pepa, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. *Overview of the CLEF-2019 CheckThat! lab: Automatic identification and verification of claims*. CheckThat! Lab (Lugano).
- Aydın, Ömer, and Enis Karaarslan. 2023. "Is ChatGPT leading generative AI? What is beyond expectations?" *Academic Platform Journal of Engineering and Smart Systems*. 11 (3): 118-134.
- Azamfirei, Razvan, Sapna R Kudchadkar, and James Fackler. 2023. "Large language models and the perils of their hallucinations." *Critical Care* 27 (1): 1-2.
- Azevedo, Pedro, Gil Rocha, Diego Esteves, and Henrique Lopes Cardoso. 2021. "Towards Better Evidence Extraction Methods for Fact-Checking Systems." *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*: 277-284.
- Babakar, Mevan, and Will Moy. 2016. "The state of automated factchecking." *Full Fact* 28.
- Balland, Pierre-Alexandre, Cristian Jara-Figueroa, Sergio G Petralia, Mathieu PA Steijn, David L Rigby, and César A Hidalgo. 2020. "Complex economic activities concentrate in large cities." *Nature human behaviour* 4 (3): 248-254.
- Baly, Ramy, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. "What was written vs. who read it: news media profiling using text analysis and social media context." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Barbizan, Silvio Nestor. 2021. "Jornalismo em vídeo gerado por inteligência artificial: narrativas e credibilidade." Doutor em Comunicação Social, Universidade Católica do Rio Grande do Sul.
- Barbrook, Richard 2009. "Imaginary Futures: From Thinking Machines." *Review of Policy Research*. 26 (3): 348-350.
- Bardoel, Jo, and Leen dHaenens. 2004. "Media responsibility and accountability. New conceptualizations and practices." *Communications* 29 (1): 5-25.
- Barker, Gary Knaul, and Susan Rich. 1992. "Influences on adolescent sexuality in Nigeria and Kenya: Findings from recent focus-group discussions." *Studies in family planning* 23 (3): 199-210.
- Barrón-Cedeño, Alberto, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, and Alex Nikolov. 2020. "Overview of CheckThat! 2020: Automatic identification and verification of claims in social media." Experimental IR Meets Multilinguality,

- Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11.
- Bateman, Jon. 2022. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace.
- Baum, William M. 2018. *Comprender o Behaviorismo: Comportamento, Cultura e Evolução*. São Paulo: Artmed Editora.
- Bayer, Justin, and Christian Osendorfer. 2015. "Learning Stochastic Recurrent Networks." Conference paper at ICLR.
- Bélanger, France, and Lemuria Carter. 2008. "Trust and risk in e-government adoption." *The Journal of Strategic Information Systems* 17 (2): 165-176.
- Bender, Emily M. 2013. "Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax." *Synthesis lectures on human language technologies* 6 (3): 1-184.
- Benko, Attila, and Cecília Sik Lányi. 2009. "History of artificial intelligence." In *Encyclopedia of Information Science and Technology, Second Edition, 1759-1762*. IGI global.
- Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. Sebastopol: O'Reilly Media, Inc.
- Birhane, Abeba. 2021. "Algorithmic injustice: a relational ethics approach." *Patterns* 2 (2): 100205.
- Blodgett-Ford, Sayoko, Thomas MS Hemnes, and Imogen Bowden. 2023. *Channeling the Explosive Power of Generative Artificial Intelligence Board oversight at the technology frontier*.
- Bonet-Jover, Alba, Alejandro Piad-Morffis, Estela Saquete, Patricio Martínez-Barco, and Miguel Ángel García-Cumbreras. 2021. "Exploiting discourse structure of traditional digital media to enhance automatic fake news detection." *Expert systems with applications* 169: 114340.
- Borin, Lars, Mia-Marie Hammarlin, Dimitrios Kokkinakis, and Fredrik Miegel. 2024. "Vaccine Hesitancy in the Nordic Countries: Trust and Distrust During the COVID-19 Pandemic." Lund: Taylor & Francis.
- Borning, Alan, Batya Friedman, Janet Davis, and Peyina Lin. 2005. "Informing public deliberation: Value sensitive design of indicators for a large-scale urban simulation." ECSCW 2005: Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work, 18–22 September 2005, Paris, France.
- Borning, Alan, Batya Friedman, and P Kahn. 2004. "Designing for human values in an urban simulation system: Value sensitive design and participatory design." Proceedings From the Eighth Biennial Participatory Design Conference.
- Borrás, Susana, and Charles Edquist. 2013. "The choice of innovation policy instruments." *Technological forecasting and social change* 80 (8): 1513-1522.
- Botnevik, Bjarte, Eirik Sakariassen, and Vinay Setty. 2020. "Brenda: Browser extension for fake news detection." Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval.
- Brand, Erik, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. 2022. "A neural model to jointly predict and explain truthfulness of statements." *ACM Journal of Data and Information Quality* 15 (1): 1-19.

- Braun, Virginia, and Victoria Clarke. 2014. "What can "thematic analysis" offer health and wellbeing researchers?" 9 (1): 261-52.
- Broinowski, Anna. 2022. "Deepfake Nightmares, Synthetic Dreams: A Review of Dystopian and Utopian Discourses Around Deepfakes, and Why the Collapse of Reality May Not Be Imminent—Yet." *Journal of Asia-Pacific Pop Culture* 7 (1): 109-139.
- Broussard, Meredith. 2015. "Artificial Intelligence for Investigative Reporting: Using an expert system to enhance journalists' ability to discover original public affairs stories." *Digital Journalism* 3 (6): 814-831.
- . 2023. *More than a glitch: Confronting race, gender, and ability bias in tech*. Cambridge: MIT Press.
- Brown, Alexander. 2018. "What is so special about online (as compared to offline) hate speech?" *Ethnicities* 18 (3): 297-326.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. "Language models are few-shot learners." *Advances in neural information processing systems* 33: 1877-1901.
- Bucchi, Massimiano. 1996. "When scientists turn to the public: Alternative routes in science communication." *Public understanding of science* 5 (4): 375.
- Bucci, Eugênio. 2018. "Pós-política e corrosão da verdade." *Revista Usp* (116): 19-30.
- Buo, Shadrack Awah. 2020. The emerging threats of deepfake attacks and countermeasures. In *arXiv preprint arXiv:2012.07989*.
- Buolamwini, Joy, and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency.
- Campbell, Donald T, and Donald W Fiske. 1959. "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological bulletin* 56 (2): 81.
- Campbell, Murray, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. "Deep blue." *Artificial intelligence* 134 (1-2): 57-83.
- Canavilhas, João, and Pollyana Ferrari. 2018. "Fact-checking: o jornalismo regressa às origens." *Jornalismo em tempo de transformação: desafios de produção e de ação*: 30-49.
- Canavilhas, João, and Thaís de Mendonça Jorge. 2022. "Fake news explosion in Portugal and Brazil the pandemic and journalists' testimonies on disinformation." *Journalism and media* 3 (1): 52-65.
- Cardoso, Gustavo. 2023. *A comunicação da comunicação. As Pessoas são a Mensagem*. Lisboa: Editora Mundos Sociais.
- Carlson, Matt. 2015. The Robotic Reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism* 3, 3 (2015), 416-431.
- Carpuat, Marine, Marie-Catherine de Marneffe, and Ivan Meza-Ruiz. 2022. "Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies." Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
- Carrol, J. M. . 1999. "Five reasons for scenario-based design." Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences., 5-8 Jan. 1999.

- Case, Donald O, J David Johnson, James E Andrews, Suzanne L Allard, and Kimberly M Kelly. 2004. "From two-step flow to the Internet: The changing array of sources for genetics information seeking." *Journal of the American Society for Information Science and Technology* 55 (8): 660-669.
- Casillas, Ramón, Helena Gómez-Adorno, Victor Lomas-Barrie, and Orlando Ramos-Flores. 2022. "Automatic Fact Checking Using an Interpretable Bert-Based Architecture on COVID-19 Claims." *Applied Sciences* 12 (20): 10644.
- Cassedy, James H. 1964. "Muckraking and Medicine: Samuel Hopkins Adams." *American Quarterly* 16 (1): 85-99.
- Castells, Manuel. 1992. *A sociedade em rede - A era da informação: economia, sociedade e cultura*. Vol. 2. *O poder da identidade*. São Paulo: Paz e Terra.
- . 2003. *A Galáxia Internet: reflexões sobre a Internet, negócios e a sociedade*. São Paulo: Zahar.
- . 2005. "A sociedade em rede: do conhecimento à política." *A sociedade em rede: do conhecimento à ação política*: 17-30.
- Cazetta, Jhonny Póvoa, and Ana Isabel Reis. 2019. "As fontes dos serviços de fact-checking luso-brasileiros." *Comunicação pública* 14 (27).
- Cenci, Alessandra, and Dylan Cawthorne. 2020. "Refining value sensitive design: A (capability-based) procedural ethics approach to technological design for well-being." *Science and Engineering Ethics* 26 (5): 2629-2662.
- Cervo, Al, Pa Bervian, and R Da Silva. 2009. *Metodologia científica*. 6. Edição. 4. Reimpressão. São Paulo: Pearson Prentice Hall.
- Chandler, Simon. 2020. "Reuters uses AI to prototype first ever automated video reports." *Forbes*.
- Charaudeau, Patrick 2008. "La médiatisation de la science. Clonage, OGM, manipulations génétiques." *Revista Latinoamericana de Estudios del Discurso* 13 (2): 121-128.
- Charles, Anderson Cordeiro, Livia Ruback, and Jonice Oliveira. 2022. "Fakepedia corpus: A flexible fake news corpus in portuguese." International Conference on Computational Processing of the Portuguese Language.
- Cheema, Gullal S, Sherzod Hakimov, and Ralph Ewerth. 2020. "Check_square at checkthat! 2020: Claim detection in social media via fusion of transformer and syntactic features." CEUR workshop proceedings, Thessaloniki, Greece.
- Chen, Zhouhan, and Juliana Freire. 2020. "Proactive discovery of fake news domains from real-time social media feeds." Companion Proceedings of the Web Conference 2020.
- Cheney-Lippold, John. 2017. "We are data." In *We Are Data*. New York: New York University Press.
- Chesney, Bobby, and Danielle Citron. 2019. "Deep fakes: A looming challenge for privacy, democracy, and national security." *Calif. L. Rev.* 107: 1753.
- Cho, Beomsang, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbbba, and Kristen Moore. 2023. "Towards understanding of deepfake videos in the wild." Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.
- Christensen, Clayton M. 1997. *The Innovator's Dilemma*. Boston: Harvard Business School Press.
- Chugh, Dolly, Max H Bazerman, and Mahzarin R Banaji. 2005. "Bounded ethicality as a psychological barrier to recognizing conflicts of interest." *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy* 17 (1): 74-95.

- "ClaimReview." 2023. <https://www.claimreviewproject.com>.
- Clarke, Victoria, and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners*. London: SAGE Publications.
- . 2017. "Thematic analysis." *The journal of positive psychology* 12 (3): 297-298.
- Clayton, Katherine, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, and Evan Morgan. 2020. "Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media." *Political behavior* 42: 1073-1095.
- Clement, Andrew. 1996. "Computing at Work: Empowering Action by." *Computerization and controversy: value conflicts and social choices*: 383.
- Clerwall, Christer. 2014. "Enter the robot journalist: Users' perceptions of automated content." *Journalism practice* 8 (5): 519-531.
- Cleveland, Keanna. 2022. "Creepy or Cool? An Exploration of Non-Malicious Deepfakes Through Analysis of Two Case Studies." University of Maryland, College Park.
- Coddington, Mark, Logan Molyneux, and Regina G Lawrence. 2014. "Fact checking the campaign: How political reporters use Twitter to set the record straight (or not)." *The International Journal of Press/Politics* 19 (4): 391-409.
- Coello-Guilarte, Laritza, Rosa María Ortega-Mendoza, Luis Villaseñor-Pineda, and Manuel Montes-y-Gómez. 2019. "Crosslingual depression detection in twitter using bilingual word alignments." *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings* 10.
- Cohn, Victor, and Lewis Cope. 2011. *News and Numbers: A Writer's Guide to Statistics*. 3 ed. Chichester: John Wiley & Sons.
- Compton, Josh. 2013. "Inoculation theory." *The SAGE handbook of persuasion: Developments in theory and practice* 2: 220-237.
- Consentino, André Tezza. 2020. "Pós-verdade e mediação política: netnografia das páginas de Facebook dos candidatos à presidência do Brasil em 2018." Diss. Universidade Fernando Pessoa.
- Cowley, Dana. 2018. "Siren at FMX 2018: Crossing the Uncanny Valley in Real Time." Siren at FMX 2018.
- Crawford, Kate. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale: Yale University Press.
- Crawford, Kate, and Trevor Paglen. 2021. "Excavating AI: The politics of images in machine learning training sets." *Ai & Society* 36 (4): 1105-1116.
- Creswell, John W, and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. 5 ed. Los Angeles: Sage publications.
- Critchley, Simon. 2011. *On humour. Thinking in Action*. Essex: Routledge.
- Cushman, Mimi, Brian S Johnson, Oliver D King, Aaron D Gitler, and James Shorter. 2010. "Prion-like disorders: blurring the divide between transmissibility and infectivity." *Journal of cell science* 123 (8): 1191-1201.
- Czeskis, Alexei, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. 2010. "Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety." *Proceedings of the sixth symposium on usable privacy and security*.
- "D-ID." 2023. Future-Facing Generative AI. <https://www.d-id.com>.

- da Silva, Mayara Karla Dantas, Maria Elizabeth Baltar Carneiro de Albuquerque, and Maria do Socorro Furtado Veloso. 2019. "Representação da informação noticiosa pelas agências de fact-checking: do acesso à informação ao excesso de informação." *Revista Brasileira de Biblioteconomia e Documentação* 15 (2): 410-426.
- da Silva Oliveira, Antonio Batista, and Paula Chimenti. 2021. "'Humanized Robots': A Proposition of Categories to Understand Virtual Influencers." *Australasian Journal of Information Systems* 25.
- Dai, Shih-Chieh, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. "Ask to Know More: Generating Counterfactual Explanations for Fake Claims." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC, USA.
- Dale, Robert. 2021. "GPT-3: What's it good for?" *Natural Language Engineering* 27 (1): 113-118.
- Darley, Andrew. 2009. *Visual digital culture: surface play and spectacle in new media genres*. London: Routledge.
- Davenport, Thomas H, and Rajeev Ronanki. 2018. "Artificial intelligence for the real world." *Harvard business review* 96 (1): 108-116.
- David, Matthew, and Carole D Sutton. 2004. *Social research: The basics*. Vol. 74. Vol. 3: Sage.
- Davis, Ernest. . 2014. "The singularity and the state of the art in artificial intelligence: The technological singularity (Ubiquity symposium)." *Ubiquity* 1-12.
- de Barros, Diana Luz Pessoa. 2020. "As fake news e as "anomalias". " *Verbum*. 9 (2): 26-41.
- De Brito Silva, Marianny Jessica, Lorena de Oliveira Ramos Delfino, Kaetana Alves Cerqueira, and Patrícia de Oliveira Campos. 2022. "Avatar marketing: a study on the engagement and authenticity of virtual influencers on Instagram." *Social Network Analysis and Mining* 12 (1): 130.
- De Caprio, Chiara, and Andrea Salvo Rossi. 2022. "Rhetorical Strategies and the Manipulation of Discourse in Machiavelli's Writings." *Past & Present* 257 (Supplement_16): 233-260.
- de Lima Gomes, Taianne, and Valquiria Aparecida Passos Kneipp. 2020. "Representatividade Racial: da tela do Jornal Nacional para discussões na Internet." *Esferas* (18): 38-38.
- DeFleur, Margaret H. 2013. *Computer-assisted investigative reporting: Development and methodology*. Routledge.
- Denning, Peter J. 2023. "Can Generative AI Bots Be Trusted?" *Communications of the ACM* 66 (6): 24-27.
- Denton, Emily L, Soumith Chintala, and Rob Fergus. 2015. "Deep generative image models using a laplacian pyramid of adversarial networks." *Advances in neural information processing systems* 28.
- Deuze, Mark. 2005. "What is journalism? Professional identity and ideology of journalists reconsidered." *Journalism* 6 (4): 442-464.
- Diakopoulos, Nicholas. 2016. "Computational journalism and the emergence of news platforms." In *The Routledge companion to digital journalism studies*, 176-184. New York: Routledge.
- . 2019. "Towards a design orientation on algorithms and automation in news production." *Digital Journalism* 7 (8): 1180-1184.

- Diakopoulos, Nicholas, and Deborah Johnson. 2021. "Anticipating and addressing the ethical implications of deepfakes in the context of elections." *New Media & Society* 23 (7): 2072-2098.
- Díaz, Patricia. 2022. "Tendencias actuales en el acceso a la información: El caso de flipboard como agregador de contenidos." *VISUAL REVIEW. International Visual Culture Review / Revista Internacional de Cultura Visual* 9: 1-22.
- Dierickx, Laurence, Carl-Gustav Lindén, and Andreas Lothe Opdahl. 2023. "The Ethical Dimensions of Data Quality for Automated Fact-Checking." *choices* 45 (51): 55.
- Dimitrakopoulou, Dimitra, and Seth C Lewis. 2023. "The generative dialogue framework and the pursuit of better listening by journalists: A design-centered approach for more constructive conversations with audiences." *Digital Journalism* 11 (3): 547-568.
- Ding, Yasan, Bin Guo, Yan Liu, Yunji Liang, Haocheng Shen, and Zhiwen Yu. 2022. "Metadetector: Meta event knowledge transfer for fake news detection." *ACM Transactions on Intelligent Systems and Technology (TIST)* 13 (6): 1-25.
- Dobbs, Michael. 2012. *The Rise of Political Fact-checking How Reagan Inspired a Journalistic Movement: A Reporter's Eye View*. New York: New America Foundation.
- Domingos, Pedro. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books.
- Doona, Joanna. 2024. "5 "Read the room Sweden"." *Vaccine Hesitancy in the Nordic Countries*: 68.
- Dourish, Paul, and Genevieve Bell. 2011. *Divining a digital future: Mess and mythology in ubiquitous computing*. Cambridge: Mit Press.
- Eco, Umberto. 1965. *Apocalípticos e integrados*. Barcelona: Editorial Lumen.
- . 2005. *Vegetal and Mineral Memory: The Future of Books: Lecture Delivered by Umberto Eco at the Bibliotheca Alexandrina on 1 November 2003*. Alexandria: Bibliotheca Alexandrina.
- . 2016. *Écrits sur la pensée au Moyen Âge*. Paris: Grasset.
- Edler, Jakob, and Jan Fagerberg. 2017. "Innovation policy: what, why, and how." *Oxford Review of Economic Policy* 33 (1): 2-23.
- "Educage Project." 2023. EducAGE. <https://educage-project.eu>.
- Ehsan, Upol, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. "The algorithmic imprint." 2022 ACM Conference on Fairness, Accountability, and Transparency.
- Elhadad, Mohamed K, Kin Fun Li, and Fayez Gebali. 2021. "COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information on COVID-19." *Advances in Intelligent Networking and Collaborative Systems: The 12th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2020)* 12.
- Elizabeth, Jane. 2014. "Who are you calling a fact checker." *Arlington: American Press Institute*.
- "Ellinika." 2019. Ellinika Hoaxes. Accessed 18 Jan. <https://www.ellinikahoaxes.gr/>
- Ellis, Emma Grey. 2018. "People can put your face on porn—and the law can't help you." *Wired*.
- Essenfelder, Renato, João Canavilhas, Haline Costa Maia, and Ricardo Jorge Pinto. 2019. "Automated News in Brazilian television: a case study on the AIDA system (Globo-Brazil)." *Doxa Comunicación* (29).

- "European Commission." 2021. COM(2021) 205 final.
https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.
- "Exorde." 2023. Clear the way in the web jungle. <https://exorde.network>.
- Eysenbach, Gunther, and James E Till. 2001. "Ethical issues in qualitative research on internet communities." *BMJ* 323 (7321): 1103-1105.
- "Fact Check Assistant." 2023. <https://factcheckassistant.socialmedialab.ca>.
- "Factiveverse." 2023. AI Editor. <https://www.factiveverse.no/ai-editor>.
- Felle, Tom. 2016. "Digital watchdogs? Data reporting and the news media's traditional 'fourth estate' function." *Journalism* 17 (1): 85-96.
- Flanagan, Mary, Daniel C Howe, and Helen Nissenbaum. 2008. "Embodying values in technology: Theory and practice." *Information Technology and Moral Philosophy*: 322-353.
- Floridi, Luciano. 2018. "Soft ethics and the governance of the digital." *Philosophy & Technology* 31: 1-8.
- Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, and Francesca Rossi. 2018. "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations." *Minds and machines* 28: 689-707.
- . 2021. "An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations." *Ethics, governance, and policies in artificial intelligence*: 19-39.
- Floridi, Luciano, Josh Cowls, Thomas C King, and Mariarosaria Taddeo. 2021. "How to design AI for social good: Seven essential factors." *Ethics, Governance, and Policies in Artificial Intelligence*: 125-151.
- Floridi, Luciano, and Mariarosaria Taddeo. 2016. "What is data ethics?" *Philosophical transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2083): 20160360-20160360.
- Folens, Tomas. 2023. "Ethical Comment on Synthesia & the Curious Case of Deepfakes." *The Elinsight*.
- Folger, Robert, and Christopher Stein. 2017. "Abduction 101: Reasoning processes to aid discovery." *Human Resource Management Review* 27 (2): 306-315.
- Følstad, Asbjørn, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, and Carolin Ischen. 2021. "Future directions for chatbot research: an interdisciplinary research agenda." *Computing* 103 (12): 2915-2942.
- Francis, Dave, and John Bessant. 2005. "Targeting innovation and implications for capability development." *Technovation* 25 (3): 171-183.
- Freier, Nathan G. 2008. "Children attribute moral standing to a personified agent." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 343-352.
- Frey, Carl Benedikt, and Michael Osborne. 2013. *The future of employment*. Oxford: Oxford Martin Programme on Technology and Employment.
- Friedman, Batya. 1997. *Human values and the design of computer technology*. Vol. 72. Cambridge: Cambridge University Press.
- Friedman, Batya, and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. Cambridge: Mit Press.

- Friedman, Batya, David G Hendry, and Alan Borning. 2017. "A survey of value sensitive design methods." *Foundations and Trends® in Human-Computer Interaction* 11 (2): 63-125.
- Friedman, Batya, Peter H Kahn Jr, Jennifer Hagman, Rachel L Severson, and Brian Gill. 2006. "The watcher and the watched: Social judgments about privacy in a public place." *Human-Computer Interaction* 21 (2): 235-272.
- Frost, Karen, Erica Frank, and Edward Maibach. 1997. "Relative risk in the news media: a quantification of misrepresentation." *American Journal of Public Health* 87 (5): 842-845.
- Furtado, Guilherme Eustáquio, Cláudia Vaz, Antonio Bovolini, Ermelinda Marques, Nuno Serra, Ana Raquel Costa-Brito, and Carolina Vila-Chã. 2022. "The impact of physical activity levels and functional fitness status on the quality of life perceived by older adults living in rural and urban areas: the Portuguese inland case." *Healthcare* 10 (7): 1266.
- "G1." 2020. Inteligência artificial: leia reportagens sobre a posse de prefeitos e vereadores em mais de 5 mil cidades, uma iniciativa inédita do G1. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://g1.globo.com/politica/noticia/2021/01/01/inteligencia-artificial-leia-reportagens-sobre-a-posse-de-prefeitos-e-veredores-em-mais-de-5-mil-cidades-uma-iniciativa-inedita-do-g1.ghtml&ved=2ahUKEwjV4_WW0bWFAxWRFRAIHfeuDfkQFnoECA8QAQ&usg=AOvVaw02ps732sEsXkgLb4FYPIlm.
- Galhardi, Cláudia Pereira, Neyson Pinheiro Freire, Maria Cecília de Souza Minayo, and Maria Clara Marques Fagundes. 2020. "Fato ou Fake? Uma análise da desinformação frente à pandemia da Covid-19 no Brasil." *Ciência & Saúde Coletiva* 25: 4201-4210.
- Gaver, Bill, Tony Dunne, and Elena Pacenti. 1999. "Design: cultural probes." *interactions* 6 (1): 21-29.
- Ge, Xun, and Susan M Land. 2003. "Scaffolding students' problem-solving processes in an ill-structured task using question prompts and peer interactions." *Educational technology research and development* 51 (1): 21-38.
- Gehlen, Marco Antônio. 2016. "Jornalismo de (im) precisão: o conhecimento matemático e a apuração de números."
- Geitgey, Adam. 2017. "Machine learning is fun part 8: How to intentionally trick neural networks." *Machine Learning is Fun*. 7 (3): 2019.
- Gencheva, Pepa, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. "A context-aware approach for detecting worth-checking claims in political debates." *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*: 267-276.
- Genesini, Silvio. 2018. "A pós-verdade é uma notícia falsa." *Revista Usp* (116): 45-58.
- Geniole, Shawn N, Brian M Bird, Alayna Witzel, Jordan T McEvoy, and Valentina Proietti. 2022. "Preliminary evidence that brief exposure to vaccination-related internet memes may influence intentions to vaccinate against COVID-19." *Computers in human behavior* 131: 107218.
- Gereffi, Gary, John Humphrey, and Timothy Sturgeon. 2005. "The governance of global value chains." *Review of international political economy* 12 (1): 78-104.
- Gibney, Elizabeth. 2017. "The scientist who spots fake videos." *Nature*.

- Giering, Maria Eduarda, and JAC de Souza. 2013. "Informar e captar: objetos de discurso em artigos de divulgação científica para crianças." *Referenciação: teoria e prática*: 205-232.
- Gillárová, Kateřina Sv, Alice N Tejkalová, and Filip Láb. 2014. "The Undressed Newsroom: The application of visual ethnography in media research." *Journalism Practice* 8 (5): 607-618.
- Goodfellow, I, Y Bengio, and A Courville. 2016. *Deep learning*. Boston: MIT Press.
- Granado, António, Dora Santos Silva, and Paulo Nuno Vicente. 2020. "Inovação nos media e nas indústrias criativas limítrofes—Uma introdução." *Media & Jornalismo* 20 (36): 5-9.
- Graves, Lucas. 2016. *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.
- Graves, Lucas, and Michelle Amazeen. 2019. "Fact-checking as idea and practice in journalism." edited by J. Nussbaum. Oxford: Oxford University Press.
- Graves, Lucas, and Federica Cherubini. 2016. *The rise of fact-checking sites in Europe*. University of Oxford (Oxford).
- Graves, Lucas, Brendan Nyhan, and Jason Reifler. 2016. "Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking." *Journal of communication* 66 (1): 102-138.
- Gray, Jonathan, Lucy Chambers, and Liliana Bounegru. 2012. *The data journalism handbook: How journalists can use data to improve the news*. Sebastopol: O'Reilly Media, Inc.
- Grigorovici, Dan M, and Corina D Constantin. 2004. "Experiencing interactive advertising beyond rich media: Impacts of ad type and presence on brand effectiveness in 3D gaming immersive virtual environments." *Journal of Interactive Advertising* 5 (1): 22-36.
- Gruzd, Anatoliy, and Philip Mai. 2020. "Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter." *Big Data & Society*. 7 (2).
- Güera, David, and Edward J Delp. 2018. "Deepfake video detection using recurrent neural networks." *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*: 1-6.
- Guerrero-Bote, Vicente P, Zaida Chinchilla-Rodríguez, Abraham Mendoza, and Félix de Moya-Anegón. 2021. "Comparative analysis of the bibliographic data sources Dimensions and Scopus: An approach at the country and institutional levels." *Frontiers in Research Metrics and Analytics* 5.
- Guest, Greg, Arwen Bunce, and Laura Johnson. 2006. "How many interviews are enough? An experiment with data saturation and variability." *Field methods* 18 (1): 59-82.
- Guimarães Rosa, João. 1986. *Grande sertão: veredas*. Rio de Janeiro: Nova Fronteira.
- Guizzo, Erico Marui. 2003. "The essential message: Claude Shannon and the making of information theory.", Massachusetts Institute of Technology.
- Güldenpfennig, Florian, Francisco Nunes, Eva Ganglbauer, and Geraldine Fitzpatrick. 2016. "Making space to engage: an open-ended exploration of technology design with older adults." *International Journal of Mobile Human Computer Interaction (IJMHCI)*. 8 (2): 1-19.
- Habermas, Jürgen. 1974. *Theory and practice*. Boston: Beacon Press.

- . 2011. "13—A publicidade como princípio de mediação entre a política e a moral". *Mudança estrutural da esfera pública. Trad. Denílson Luiz Werle. São Paulo: Editora da UNESP: 264-287.*
- Hafez, Kai. 2002. "Journalism ethics revisited: A comparison of ethics codes in Europe, North Africa, the Middle East, and Muslim Asia." *Political communication* 19 (2): 225-250.
- Haguette, Teresa Maria Frota. 2013. "Metodologias qualitativas na sociologia." In *Metodologias qualitativas na sociologia.*, 223-223.
- Hall, Justin R, Selen Savas-Hall, and Eric H Shaw. 2022. "A deductive approach to a systematic review of entrepreneurship literature." *Management Review Quarterly*: 1-30.
- "Hamilton 2.0." 2023. You Can Help Secure Democracy. <https://securingdemocracy.gmfus.org/about-us/>.
- Hammouchi, Hicham, and Mounir Ghogho. 2022. "Evidence-aware multilingual fake news detection." *Ieee Access* 10: 116808-116818.
- Hancock, Dawson R, Bob Algozzine, and Jae Hoon Lim. 2021. *Doing case study research: A practical guide for beginning researchers*. Vol. 10. New York: Teacher College Press.
- Harrag, Fouzi, and Mohamed Khalil Djahli. 2022. "Arabic fake news detection: A fact checking based deep learning approach." *Transactions on Asian and Low-Resource Language Information Processing*. 21 (4): 1-34.
- Harrer, Stefan. 2023. "Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine." *EBioMedicine* 90.
- Hassan, Naeemul, Chengkai Li, and Mark Tremayne. 2015. "Detecting check-worthy factual claims in presidential debates." Proceedings of the 24th acm international on conference on information and knowledge management.
- Helberger, Natali, Max van Drunen, Judith Moeller, Sanne Vrijenhoek, and Sarah Eskens. 2022. "Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals." 10 (10): 1605-1626.
- Heravi, Bahareh Rahmanzadeh, and Jarred McGinnis. 2015. "Introducing social semantic journalism." *The Journal of Media Innovations* 2 (1): 131-140.
- Hesse, Bradford W, Alexandra J Greenberg, and Lila J Finney Rutten. 2016. "The role of Internet resources in clinical oncology: promises and challenges." *Nature Reviews Clinical Oncology* 13 (12): 767-776.
- Hidalgo, César A, Diana Orghian, Jordi Albo Canals, Filipa de Almeida, and Natalia Martin. 2020. "Executive Summary: How Humans Judge Machines." Boston: MIT Press.
- High, Rob. 2012. "The era of cognitive systems: An inside look at IBM Watson and how it works." *IBM Corporation, Redbooks* 1: 16.
- Hilgartner, Stephen. 1990. "The dominant view of popularization: Conceptual problems, political uses." *Social studies of science* 20 (3): 519-539.
- Hill, Hamlin. 2010. *Mark Twain: God's Fool*. Chicago: University of Chicago Press.
- Hiort, A. 2022. "How many virtual influencers are there." *Virtual Humans*, Last accessed 01st December.
- Hodgetts, Darrin, Kerry Chamberlain, Margaret Scammell, Rolinda Karapu, and Linda Waimarie Nikora. 2008. "Constructing health news: possibilities for a civic-oriented journalism." *Health: 12* (1): 43-66.

- Hoffman, Jesse, Peter Pelzer, Loes Albert, Tine Béneker, Maarten Hajer, and Astrid Mangnus. 2021. "A futuring approach to teaching wicked problems." *Journal of Geography in Higher Education* 45 (4): 576-593.
- Hopkins, Bruce. 2024. "Multimodal AI: Creating a Podcast Visualizer with Whisper and DALL·E 3." In *ChatGPT for Java: A Hands-on Developer's Guide to ChatGPT and Open AI APIs.*, 99-140. New York: Springer.
- Howard, Philip N, Samuel Woolley, and Ryan Calo. 2018. "Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration." *Journal of information technology & politics* 15 (2): 81-93.
- Humprecht, Edda. 2020. "How do they debunk "fake news"? A cross-national comparison of transparency in fact checks." *Digital journalism* 8 (3): 310-327.
- Hunt, Shelby D, and Scott Vitell. 1986. "A general theory of marketing ethics." *Journal of macromarketing* 6 (1): 5-16.
- Hutchinson, Hilary, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, and Heiko Hansen. 2003. "Technology probes: inspiring design for and with families." Proceedings of the SIGCHI conference on Human factors in computing systems.
- Hwang, Yoori, Ji Youn Ryu, and Se-Hoon Jeong. 2021. "Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education." *Cyberpsychology, behavior and social networking*. 24 (3): 188-193.
- Ivancevich, Alyssa. 2022. "Deepfake Reckoning: Adapting Modern First Amendment Doctrine to Protect Against the Threat Posed to Democracy." *Hastings Const.* 49: 61.
- Jackson, Brooks. 2012. *FactCheck.org Report*.
- Jenkins, Henry. 2011. "Convergence Culture. Where Old and New Media Collide." *Revista Austral de Ciencias Sociales* 20: 129-133.
- Jensen, Olaf P, Jason S Link, Anthony DM Smith, Trevor A Branch, Selina S Heppell, Michael C Melnychuk, Steven JD Martell, John G Pope, Timothy E Essington, and Ana M Parma. 2012. "Can catch share fisheries better track management targets?" 13 (3): 267-290.
- Jerónimo, Pedro, Giovanni Ramos, and Luisa Torre. 2022. "News Deserts Europe 2022: Portugal Report." *MediaTrust*.
- Jiang, Shan, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020. "Factoring fact-checks: Structured information extraction from fact-checking articles." Proceedings of the Web Conference 2020.
- Jo, A. 2023. "The promise and peril of generative AI." *Nature* 614 (1): 214-216.
- Johnson, Peter, and Pamela Robinson. 2014. "Civic hackathons: Innovation, procurement, or civic engagement?" *Review of policy research* 31 (4): 349-357.
- JohnSon, Sally, and Astrid Ensslin. 2006. "Language in the news: Some reflections on keyword analysis using WordSmith Tools and the BNC." *Leeds Working Papers in Linguistics and Phonetics* 11: 96-109.
- Jones, Bronwyn, Ewa Luger, and Rhianne Jones. 2023. "Future Thinking with Journalists." Proceedings with Computation + Journalism European Data & Computational Journalism Conference 2023, Zurich, Switzerland.
- Jones, Marie D. 2021. *Disinformation and You: Identify Propaganda and Manipulation*. Canton: Visible Ink Press.

- Jones, Thomas M. 1991. "Ethical decision making by individuals in organizations: An issue-contingent model." *Academy of management review*. 16 (2): 366-395.
- Jurafsky, Dan, and Christopher Manning. 2012. "Natural language processing." *Instructor*. 212 (998): 3482.
- Kahan, Dan M, Kathleen Hall Jamieson, Asheley Landrum, and Kenneth Winnege. 2017. "Culturally antagonistic memes and the Zika virus: An experimental test." *Journal of Risk Research*. 20 (1): 1-40.
- Karagiannis, Georgios, Immanuel Trummer, Saehan Jo, Shubham Khandelwal, Xuezhi Wang, and Cong Yu. 2019. "Mining an" anti-knowledge base" from Wikipedia updates with applications to fact checking and beyond." *Proceedings of the VLDB Endowment* 13 (4): 561-573.
- Kaufman, Dora. 2021. "Inteligência Artificial e os desafios éticos: a restrita aplicabilidade dos princípios gerais para nortear o ecossistema de IA." *Paulius: Revista de Comunicação da Fapcom*. 5 (9).
- Kelly, Kevin. 2010. *What Technology Wants*. Kindle Edition. London: Penguin Publishing Group.
- Kerlinger, Fred Nichols. 1979. "Behavioral research: A conceptual approach." *Journal of Music Therapy*. 25 (2): 112-114.
- Keyes, Ralph. 2004. *The post-truth era: Dishonesty and deception in contemporary life*. New York: St. Martin's Press.
- Kickbusch, Ilona. 1997. "Think health: What Makes the Difference?" *Health Promotion International* 12 (4): 265-272.
- Kietzmann, Jan, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020. "Deepfakes: Trick or treat?" *Business Horizons* 63 (2): 135-146.
- Kitamura, Felipe C. 2023. "ChatGPT Is Shaping the Future of Medical Writing but Still Requires Human Judgment." 307 (2): 230171.
- Kitchenham, Barbara, and Stuart Charters. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Durham: University of Durham.
- Kitchenham, Barbara, and Shari Lawrence Pfleeger. 2002. "Principles of survey research part 4: questionnaire evaluation." *ACM SIGSOFT Software Engineering Notes*. 27 (3): 20-23.
- Kitchens, Brent, Christopher A Harle, and Shengli Li. 2014. "Quality of health-related online search results." *Decision Support Systems*. 57: 454-462.
- Koch, Bernard, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. "Reduced, reused and recycled: The life of a dataset in machine learning research." *Proceedings with NeurIPS Dataset & Benchmark track*.
- Koroteev, MV. 2021. "BERT: a review of applications in natural language processing and understanding." *arXiv preprint arXiv:2103.11943*.
- Kotonya, Neema, and Francesca Toni. 2020. "Explainable automated fact-checking for public health claims." *arXiv preprint arXiv:2010.09926*.
- Krueger, Richard A. 2014. *Focus groups: A practical guide for applied research*. 5 ed. Washington DC: Sage publications.
- Kulkarni, Akshay, Adarsha Shivananda, Anoosh Kulkarni, and Dilip Gudivada. 2023. "Introduction to Generative AI." In *Applied Generative AI for Beginners: Practical Knowledge on Diffusion Models, ChatGPT, and Other LLMs*, 1-13. Springer.
- Kvale, Steinar. 1996. "Interviews: an introduction to qualitative research interviewing/Steinar Kvale." *American Journal of Evaluation*: 326.

- Laatsit, Mart, Markus Grillitsch, and Lea Fünfschilling. 2022. *Great expectations: The promises and limits of innovation policy in addressing societal challenges*. Lund University, CIRCLE-Centre for Innovation Research.
- Lab, Ryerson University Social Media. 2020. COVID-19 Fact-checkers Dataset. edited by Canadian Institutes of Health Research (CHIR): 2019 Novel Coronavirus Rapid Research.
- Lahoti, Pragya, and Suneel Prasad. 2022. "Using machine learning and robotic process automation for misinformation detection on Twitter: Analysing the tweets on Covid-19 pandemic." AIP Conference Proceedings.
- Lamprou, Evangelos, Nikos Antonopoulos, Iouliani Anomeritou, and Chrysoula Apostolou. 2021. "Characteristics of fake news and misinformation in greece: the rise of new crowdsourcing-based journalistic fact-checking models." *Journalism and Media* 2 (3): 417-439.
- Lanius, Candice, Ryan Weber, and William I MacKenzie Jr. 2021. "Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey." *Social network analysis and mining* 11 (1): 32.
- Latoschik, Marc Erich, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. 2017. "The effect of avatar realism in immersive social virtual realities." Proceedings of the 23rd ACM symposium on virtual reality software and technology.
- Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, and David Rothschild. 2018. "The science of fake news." *Science* 359 (6380): 1094-1096.
- LeCun, Y. 2016. "The next frontier in AI: Unsupervised learning" <https://www.youtube.com/watch?v=IbjF5VjniVE>.
- Lemos, A. 2013. "A Comunicação das coisas: Teoria ator-rede e cibercultura." *Aborto no Brasil e Países do Cone Sul: Panorama da situação e dos estudos acadêmicos*: 207-226.
- Leonardi, Paul M. 2013. "Theoretical foundations for the study of sociomateriality." *Information and organization* 23 (2): 59-76.
- Letham, Benjamin, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model." *Ann. Appl. Stat.* 9 (3): 1350-1371.
- Lévy, Pierre. 1997. "Cyberculture." *Rapport au Conseil de l'Europe*. Odile Jacob.
- . 1998. "A inteligência coletiva." *São Paulo: Loyola* 22.
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33: 9459-9474.
- Lima, Helena, and Jorge Pedro Sousa. 2021. "Enquadramentos noticiosos da Gripe Pneumónica na imprensa portuguesa: os casos dos jornais O Comércio do Porto, O Algarve e da revista Ilustração Portuguesa." In *Para uma história do jornalismo em Portugal.*, 203-237.
- Lin, Zhaojiang, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. "Caire: An end-to-end empathetic chatbot." Proceedings of the AAAI Conference on Artificial Intelligence.
- Lippmann, Walter. 2017. *The phantom public*. New York: Routledge.

- Liu, Kunlin, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, and Jian Jiang. 2020. "DeepFaceLab: Integrated, flexible and extensible face-swapping framework." *arXiv preprint arXiv:2005.05535* 141.
- Liu, Lihui, Boxin Du, Jiejun Xu, Yinglong Xia, and Hanghang Tong. 2022. "Joint knowledge graph completion and question answering." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
- Liu, Yixin, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, and Jianfeng Gao. 2024. "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models." *arXiv preprint arXiv:2402.17177*.
- Locke, Karen, Martha S Feldman, and Karen Golden-Biddle. 2015. "Discovery, validation, and live coding." *Handbook of qualitative organizational research: Innovative pathways and methods*: 371-380.
- Lundqvist, Martin. 2019. "Post-war memorialisation as everyday peace? Exploring everyday (dis-) engagements with the Maoist martyrs' gate of Beni Bazaar in Nepal." *Conflict, Security & Development* 19 (5): 475-496.
- "Macedo De Cavaleiros." 2024. Câmara Municipal - Universidade Sénior. Accessed 5 de abril. <https://www.cm-macedodecavaleiros.pt/pages/262>.
- Macefield, Ritch. 2009. "How to specify the participant group size for usability studies: a practitioner's guide." *Journal of usability studies* 5 (1): 34-45.
- Macefield, Ritchie. 2006. "Conceptual models and usability." In *Encyclopedia of Human Computer Interaction*, 112-119. Hershey: IGI Global.
- Maia, Haline. 2024. Artificial Intelligence and Infodemic: Video Dataset for Fact-Checked Health Communication and Synthetic Media [Data set]. INESC TEC.
- Maksl, Adam, Seth Ashley, and Stephanie Craft. 2015. "Measuring news media literacy." *Journal of Media Literacy Education* 6 (3): 29-45.
- Mansour, Watheq, Tamer Elsayed, and Abdulaziz Al-Ali. 2022. "Did i see it before? detecting previously-checked claims over twitter." European conference on information retrieval.
- Mantzaris, Alexios. 2018. "Fact-checking 101." *Journalism, fake news & disinformation: Handbook for journalism education and training*.: 85-100.
- Marcondes Filho, Ciro. 2019. "Fake news: o buraco é muito mais em baixo." Proceedings in As fake news e a nova ordem (des) informativa na era da pós-verdade.
- Marconi, F, and A Siegmán. 2017. "The Future of Augmented Journalism: A guide for Newsrooms in the age of smart machines." *About AP Report*.
- Marconi, Marina de Andrade, and Eva Maria Lakatos. 2001. *Metodologia do trabalho científico*. Vol. 6. São Paulo: Atlas.
- Marra, Francesco, Diego Gagnaniello, Davide Cozzolino, and Luisa Verdoliva. 2018. "Detection of gan-generated fake images over social networks." IEEE conference on multimedia information processing and retrieval (MIPR).
- Marres, Noortje. 2018. "Why we can't have our facts back." *Engaging Science, Technology, and Society* 4: 423-443.
- Martín, Alejandro, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2022. "FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference." *Knowledge-based systems* 251: 109265.

- Martinez-Plumed, Fernando, Bao Sheng Loe, Peter Flach, Seán O hEigeartaigh, Karina Vold, and José Hernández-Orallo. 2018. "The facets of artificial intelligence: A framework to track the evolution of AI." Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).
- Marwick, Alice E. 2013. "Why do people share fake news? A sociotechnical model of media effects." *Georgetown law technology review* 2 (2): 474-512.
- McBride, Kelly, and Tom Rosenstiel. 2013. *The new ethics of journalism: Principles for the 21st century*. London: CQ Press.
- McCarthy, John. 2007. "What is artificial intelligence." Stanford University. <http://www-formal.stanford.edu/jmc/index.html>.
- McCoy, John L. 1969. "Rokeach." Beliefs, Attitudes and Values: A Theory of Organization and Change" (Book Review). *Rural Sociology* 34 (4): 597.
- McGuffie, Kris, and Alex Newhouse. 2020. *The radicalization risks of GPT-3 and advanced neural language models*. Middlebury Institute for International Studies.
- McLuhan, Marshall. 1969. *Mutations*. Paris: Name.
- . 1971. *The Gutenberg galaxy: The making of typographic man*, Repr. London: Routledge & Kegan Paul.
- McNaughton, David, Janice Light, David R Beukelman, Chris Klein, Dana Nieder, and Godfrey Nazareth. 2019. "Building capacity in AAC: A person-centred approach to supporting participation by people with complex communication needs." *Augmentative and Alternative Communication* 35 (1): 56-68.
- "Media Bias Fact Check." 2019. The Latest Fact Checks curated. <https://mediabiasfactcheck.com/>.
- Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. 2010. "Twitter under crisis: Can we trust what we RT?" Proceedings of the first workshop on social media analytics.
- Mergulhão, Drds, M Júnior, JC Oliveira, and PF Albuquerque. 2018. "Post-truth, fake news e processo eleitoral." *Revista de Estudos Eleitorais*. 2 (4).
- Meyer, Philip. 1991. *The new precision journalism*. Bloomington: Indiana University Press.
- Meyer, Robinson. 2018. "The grim conclusions of the largest-ever study of fake news." *The Atlantic* 8: 2018.
- Miller, Ielleen. 2020. *Research Guides: Journalism - Fact-Checking Sites*.
- Millière, Raphaël. 2022. "Deep learning and synthetic media." *Synthese* 200 (3): 231.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model cards for model reporting." Proceedings of the conference on fairness, accountability, and transparency.
- Mittelstadt, Brent. 2019. "Principles alone cannot guarantee ethical AI." *Nature machine intelligence* 1 (11): 501-507.
- Moffitt, JD, Catherine King, and Kathleen M Carley. 2021. "Hunting conspiracy theories during the COVID-19 pandemic." *Social Media+ Society* 7 (3).
- Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and the PRISMA Group*. 2009a. "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." *Annals of internal medicine* 151 (4): 264-269.
- . 2009b. "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." *Annals of internal medicine* 151 (4): 264-269.
- Mojtahed, Reza, Miguel Baptista Nunes, Jorge Tiago Martins, and Alex Peng. 2014. "Equipping the constructivist researcher: The combined use of semi-structured

- interviews and decision-making maps." *Electronic Journal of Business Research Methods* 12 (2): 87-95.
- Mooney, Peter, and Levente Juhász. 2020. "Mapping COVID-19: How web-based maps contribute to the infodemic." *Dialogues in Human Geography* 10 (2): 265-270.
- Moravec, Patricia L, Antino Kim, and Alan R Dennis. 2020. "Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media." *Information Systems Research* 31 (3): 987-1006.
- Moreno-Almeida, Cristina 2024. *Memes, Monsters, and the Digital Grotesque*. Great Britain: Oxford University Press.
- Mori, Masahiro, Karl F MacDorman, and Norri Kageki. 2012. "The uncanny valley [from the field]." *IEEE Robotics & automation magazine* 19 (2): 98-100.
- Morozov, Evgeny. 2018. *Big tech*. São Paulo: Ubu Editora.
- Mrowczynski, Rafael; Witzel, Andreas; Reiter, Herwig 2013. "The Problem-Centered Interview: Principles and Practice." *Laboratorium*. 5 (2): 226-227.
- Mu, Yida, Ye Jiang, Freddy Heppell, Iknor Singh, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. "A Large-Scale Comparative Study of Accurate COVID-19 Information versus Misinformation." *arXiv preprint arXiv:2304.04811*.
- Muller, Michael J. 1999. "Invisible work of telephone operators: An ethnocritical analysis." *Computer Supported Cooperative Work (CSCW)* 8 (1-2): 31-61.
- Muller, Michael J, and Allison Druin. 2012. "Participatory design: The third space in human-computer interaction." In *Human Computer Interaction Handbook*, 1125-1153. Boca Raton: CRC Press.
- Murray, Kristen A. 2016. "The Last Laugh: Dark Comedy on US Television." *Taboo Comedy: Television and Controversial Humour*: 41-59.
- Nabożny, Aleksandra, Bartłomiej Balcerzak, Mikołaj Morzy, and Adam Wierzbicki. 2021. "Focus on misinformation: Improving medical experts' efficiency of misinformation detection." Proceedings at International Conference on Web Information Systems Engineering.
- Nakamoto, Kazuo. 2009. *Infrared and Raman spectra of inorganic and coordination compounds, part B: applications in coordination, organometallic, and bioinorganic chemistry*. 6 ed. New Jersey: John Wiley & Sons.
- Nakov, Preslav, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. "Automated fact-checking for assisting human fact-checkers." Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track.
- Nathan, Lisa P, Batya Friedman, Predrag Klasnja, Shaun K Kane, and Jessica K Miller. 2008. "Envisioning systemic effects on persons and society throughout interactive system design." Proceedings of the 7th ACM conference on Designing interactive systems.
- Neumann, Terrence, Maria De-Arteaga, and Sina Fazelpour. 2022. "Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.
- Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2018. *Digital News Report 2018*.

- Ng, Lynnette Hui Xian, and Kathleen M Carley. 2021. "'The coronavirus is a bioweapon': classifying coronavirus stories on fact-checking sites." *Computational and Mathematical Organization Theory* 27 (2): 179-194.
- Nguyen, Van-Hoang, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. "Fang: Leveraging social context for fake news detection using graph representation." Proceedings of the 29th ACM international conference on information & knowledge management.
- Nielsen, Jakob. 1994. *Usability engineering*. Mountain View: Morgan Kaufmann.
- Nielsen, Jakob, and Thomas K Landauer. 1993. "A mathematical model of the finding of usability problems." Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems.
- Nikolov, Alex, Giovanni Da San Martino, Ivan Koychev, and Preslav Nakov. 2020. "Team alex at clef checkthat! 2020: Identifying check-worthy tweets with transformer models." Proceedings at CLEF 2020.
- Nilsson, Nils Johan. 1998. *Artificial intelligence: a new synthesis*. Burlington.: Morgan Kaufmann.
- Nilsson, NJ. 2014. *Principles of artificial intelligence*. Burlington.: Morgan Kaufmann.
- Noble, Safiya Umoja. 2018. "Algorithms of oppression." In *Algorithms of oppression*. New York.: New York University Press.
- Norvig, P Russel. 2002. "ArtificialIntelligence: A modern approach." *Knowledge-Based Systems*. 90 (An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage.): 33-48.
- Nunes, Sérgio, Alípio Mario Jorge, Evelin Amorim, Hugo Sousa, António Leal, Purificação Moura Silvano, Inês Cantante, and Ricardo Campos. 2024. "Text2Story Lusa: A Dataset for Narrative Analysis in European Portuguese News Articles." Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).
- Nutbeam, Don. 2000. "Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century." *Health promotion international* 15 (3): 259-267.
- Nygren, Thomas, Mona Guath, Carl-Anton Werner Axelsson, and Divina Frau-Meigs. 2021. "Combating visual fake news with a professional fact-checking tool in education in France, Romania, Spain and Sweden." *Information* 12 (5): 201.
- Olshausen, Bruno A, and David J Field. 1996. "Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images." *Nature* 381: 607-609.
- Omezi, Natasha, and Hamid Jahankhani. 2020. "Proposed forensic guidelines for the investigation of fake news." *Policing in the Era of AI and Smart Societies*: 231-265.
- "OpenAI." 2024. How openai is approaching 2024 worldwide elections. <https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections>.
- Oshikawa, Ray, Jing Qian, and William Yang Wang. 2020. "A survey on natural language processing for fake news detection." Proceedings of the 12th Conference on Language Resources and Evaluation.
- Owen, Diana. 2017. *The state of technology in global newsrooms*. Washington, D.C.: International Center for Journalists.
- Paimre, Marianne, Sirje Virkus, and Kairi Osula. 2023. "Health information behavior and related factors among Estonians aged ≥ 50 years during the COVID-19 pandemic." *Journal of Documentation* 79 (5): 1164-1181.

- Paisana, Miguel, Ana Pinto-Martinho, and Gustavo Cardoso. 2020. "Trust and fake news: Exploratory analysis of the impact of news literacy on the relationship with news content in Portugal." *Communication & Society* 32 (2): 105-117.
- Palazzo, Guido, Franciska Krings, and Ulrich Hoffrage. 2012. "Ethical blindness." *Journal of business ethics* 109: 323-338.
- Pathak, Archita, Rohini K Srihari, and Nihit Natu. 2021. "Disinformation: analysis and identification." *Computational and Mathematical Organization Theory* 27 (3): 357-375.
- Pavlik, John, Everette Dennis, Rachel Davis Mersey, and Justin Gengler. 2018. *Mobile Disruptions in the Middle East: Lessons from Qatar and the Arabian Gulf Region in mobile media content innovation*. London: Routledge.
- Pearl, Judea, and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. New York: Basic books.
- Peeters, Marieke MM, Jurriaan van Diggelen, Karel Van Den Bosch, Adelbert Bronkhorst, Mark A Neerincx, Jan Maarten Schraagen, and Stephan Raaijmakers. 2021. "Hybrid collective intelligence in a human–AI society." *AI & society* 36 (1): 217-238.
- Pennycook, Gordon, Tyrone D Cannon, and David G Rand. 2018. "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general* 147 (12): 1865.
- Pereira, Roberto, and Maria Cecília Calani Baranauskas. 2014. "Value pie: a culturally informed conceptual scheme for understanding values in design." *Human-Computer Interaction. Theories, Methods, and Tools: 16th International Conference, HCI International., Crete, Greece*.
- Pereira, Susana, José Azevedo, and António Machiavelo. 2015. "A matemática na imprensa diária portuguesa." *Proceedings at Literacia, Media e Cidadania-Livro de Atas do 3.º Congresso*.
- Perloff, Richard M. 1993. *The dynamics of persuasion: Communication and attitudes in the 21st century*. New York: Routledge.
- Phillips, Whitney, and Ryan M Milner. 2017. "Decoding memes: Barthes' punctum, feminist standpoint theory, and the political significance of # YesAllWomen." *Entertainment values: How do we assess entertainment and why does it matter?:* 195-211.
- Pickard, Victor. 2018. "Media activism from above and below: Lessons from the 1940s American reform movement." In *Media, Ideology and Hegemony*, 229-247. Brill.
- Pinker, Steven. 2003. *The blank slate: The modern denial of human nature*. New York: Penguin.
- Pogorelov, Konstantin, Daniel Thilo Schroeder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkukova, and Johannes Langguth. 2020. "FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020." *Proceedings of the MediaEval*.
- Pogorelov, Konstantin, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. "Wico text: a labeled dataset of conspiracy theory and 5g-corona misinformation tweets." *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*.
- Poncette, Akira-Sebastian, Pablo-David Rojas, Joscha Hofferbert, Alvaro Valera Sosa, Felix Balzer, and Katarina Braune. 2020. "Hackathons as stepping stones in health care innovation: case study with systematic recommendations." *Journal of medical Internet research* 22 (3): e17004.

- Portugal, R, B Wilczek, M Eder, N Thurman, and M Haim. 2023. "Design Thinking for Journalism in the AI age: Towards an Innovation Process for Responsible AI Applications." Paper presented at the The Joint Computation + Journalism European Data & Computational Journalism Conference, Zurich.
- Posetti, Julie, and Kalina Bontcheva. 2020. *Disinfodemic*. In *Deciphering Covid-19 disinformation, Policy brief*.
- . 2021. "Infodemic: disinformation and media literacy in the context of COVID-19." *Internet Sectoral Overview 3 (13)*: 1-21.
- Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." Proceedings of the 2020 conference on fairness, accountability, and transparency.
- Rahaman, Md, MM Ahsan, Nishath Anjum, Md Rahman, and Md Nafizur Rahman. 2023a. "The AI Race is on! Google's Bard and OpenAI's ChatGPT Head to Head: An Opinion Article." *Mizanur and Rahman, Md Nafizur, The AI Race is on*.
- . 2023b. "The AI Race is on! Google's Bard and OpenAI's ChatGPT Head to Head: An Opinion Article." *Mizanur and Rahman, Md Nafizur, The AI Race is on*.
- Rakotoson, Loïc, Charles Letaillieur, Sylvain Massip, and Fréjus AA Laleye. 2022. "Extractive-boolean question answering for scientific fact checking." Proceedings of the 1st International Workshop on Multimedia AI against Disinformation.
- Rapoport, Hillel. 2016. "Migration and globalization: what's in it for developing countries?" *International Journal of Manpower 37 (7)*: 1209-1226.
- Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. "Truth of varying shades: Analyzing language in fake news and political fact-checking." Proceedings of the 2017 conference on empirical methods in natural language processing.
- Raskin, Jef. 1994. "Intuitive equals familiar." *Communications of the ACM 37 (9)*: 17-19.
- Reis, Julio CS, and Fabrício Benevenuto. 2021. "Supervised learning for misinformation detection in whatsapp." Proceedings of the Brazilian Symposium on Multimedia and the Web.
- "Reporters Lab." 2022. Fact-checkers extend their global reach with 391 outlets, but growth has slowed. Reporters Lab. Accessed January 17. <https://reporterslab.org/fact-checkers-extend-their-global-reach-with-391-outlets-but-growth-has-slowed/>.
- ResnickP, VarianHR. 1997. "Recommendersystems." *Communications of the ACM 40 (3)*: 56-58.
- Rittel, Horst WJ, and Melvin M Webber. 1973. "Dilemmas in a general theory of planning." *Policy sciences 4 (2)*: 155-169.
- Roets, Arne. 2017. "'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions." *Intelligence 65*: 107-110.
- Rokeach, Milton. 1973. *The nature of human values*.: Free press.
- Roose, Kevin. 2018. "Here come the fake videos, too." *The New York Times*.
- Rozenbeek, Jon, and Sander Van Der Linden. 2021. *Inoculation theory and misinformation*.
- Rosa, Hartmut. 2019. *Resonance: A sociology of our relationship to the world*: John Wiley & Sons.
- Roser, M, and E Ortiz-Ospina. 2016. "Global Education." *Our world in data*., 2016.

- Roslyng, Mette Marie, and Gorm Larsen. 2021. "The struggle between the power of language and the language of power: Pro-and anti-vaccination memes and the discursive construction of knowledge." *Journal of Language and Politics* 20 (6): 894-912.
- Rossmann, Gretchen B, and Sharon F Rallis. 2010. "Everyday ethics: Reflections on practice." *International Journal of Qualitative Studies in Education* 23 (4): 379-391.
- Royakkers, Lambèr, Jelte Timmer, Linda Kool, and Rinie Van Est. 2018. "Societal and ethical issues of digitization." *Ethics and Information Technology* 20: 127-142.
- Rüdiger, Francisco. 2016. "O Ocidente e a técnica: estágios reflexivos do pensamento tecnológico." *As teorias da cibercultura: perspectivas, questões e autores. Porto Alegre: Sulina: 73-104.*
- Ruiz, María José Ufarte, and Francisco José Murcia Verdú. 2018. "El fact checking: en busca de un nuevo modelo de negocio sostenible para el periodismo. Estudio de caso de Miniver." *Miguel Hernández communication journal* (9): 511-534.
- Russell, Stuart. 2021. *Inteligência artificial a nosso favor: Como manter o controle sobre a tecnologia. Companhia das Letras. São Paulo.*
- Russell, Stuart 2022. "Provably Beneficial Artificial Intelligence." 27th International Conference on Intelligent User Interfaces, Helsinki, Finland.
- Sackett, David L, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't.: British Medical Journal Publishing Group.
- Saeed, Mohammed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. "Crowdsourced fact-checking at Twitter: How does the crowd compare with experts?" Proceedings of the 31st ACM International Conference on Information & Knowledge Management.
- Sanders, Elizabeth B-N, and Pieter Jan Stappers. 2008. "Co-creation and the new landscapes of design." *Co-design* 4 (1): 5-18.
- Sanders, Elizabeth B-N, and Bo Westerlund. 2011. "Experiencing, exploring and experimenting in and with co-design spaces." *Nordes* (4).
- Santaella, Lucia. 2001. *Matrizes da linguagem e pensamento: sonora, visual, verbal: aplicações na hipermídia.* São Paulo: Editora Iluminuras Ltda.
- Saquete, Estela, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. "Fighting post-truth using natural language processing: A review and open challenges." *Expert systems with applications* 141: 112943.
- Sarrouti, Mourad, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. "Evidence-based fact-checking of health-related claims." Findings of the Association for Computational Linguistics: EMNLP 2021.
- Savage, Neil. 2021. "Fact-finding mission." *Communications of the ACM* 64 (3): 18-19.
- Sawczyn, Albert, Jakub Binkowski, Denis Janiak, Łukasz Augustyniak, and Tomasz Kajdanowicz. 2021. "Fact-checking: Relevance assessment of references in the Polish political domain." *Procedia Computer Science* 192: 1285-1293.
- Schick, Timo, and Hinrich Schütze. 2020. "Exploiting cloze questions for few shot text classification and natural language inference." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.

- Schilling, Jim, Sujeiry Guzman, Chad Lewis, and Jupil Ko. 2020. "A Comparison of the Immediate Effect of Swedish Massage vs Passive Stretching on Range of Motion in Overhead Athletes." *The Asian Journal of Kinesiology* 22 (3).
- Schuff, Hendrik, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. "Human interpretation of saliency-based explanation over text." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.
- Schwartz, Shalom H. 2012. "An overview of the Schwartz theory of basic values." *Online readings in Psychology and Culture*. 2 (1): 2307-0919.1116.
- Seaton, Jean, Amy Sippitt, and Ben Worthy. 2020. "Fact checking and information in the age of covid." *The Political Quarterly* 91 (3): 578-584.
- Shaar, Shaden, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. "Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection." Fourth NAACL 2021 Workshop on Natural Language Processing for Internet Freedom (NLP4IF) Workshop: Censorship, Disinformation, and Propaganda.
- Shao, Chengcheng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. "Hoaxy: A platform for tracking online misinformation." Proceedings of the 25th international conference companion on world wide web.
- Sharma, Deepak, and Jitendra Sharma. 2023. "The potential of virtual cloud character creation technology-meta human creator: A review." AIP Conference Proceedings.
- Sharma, Dilip Kumar, and Sonal Garg. 2021. "A Framework for Automatic Fake Content Identification." 2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM).
- Shifman, Limor. 2013. *Memes in digital culture*. Cambridge: MIT press.
- Siegel, Lee. 2008. *Against the machine: Being human in the age of the electronic mob*. New York: Random House.
- Silva, Dora Santos. 2021. "Innovation in european journalism. The case of cultural journalism." *Livros ICNOVA*.
- Silva, Elsa Costa. 2013. "Regulando pluralismo e diversidade em Portugal: uma equação sem cidadãos." *Observatorio (OBS*)* 7 (1).
- Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. 2016. "Mastering the game of Go with deep neural networks and tree search." *nature* 529 (7587): 484-489.
- Silverblatt, Art. 2004. "Media as social institution." *American Behavioral Scientist* 48 (1): 35-41.
- Simondon, Gilbert. 1989. *Du mode d'existence des objets techniques*. Paris: Editions Aubier.
- Simpson, Tyler, and Feng-Jen Yang. 2022. "Some Hands-on Approaches to Fake Political News Detection." Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning.
- Singer, Jane B. 2023. "Closing the barn door? Fact-checkers as retroactive gatekeepers of the Covid-19 "infodemic"." *Journalism & Mass Communication Quarterly* 100 (2): 332-353.
- Sippitt, Amy. 2019. "Full fact." In *Data in Society*, 359-364. Policy Press.

- Siricharoen, Waralak Vongdoiwang. 2021. "Using empathy mapping in design thinking process for personas discovering." Proceedings of the 9th EAI International Conference, ICCASA., Vietnam.
- Smeros, Panayiotis, Carlos Castillo, and Karl Aberer. 2021. "Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking." Proceedings of the 30th ACM international conference on information & knowledge management.
- Smith, Jessie J. 2022. "Developing a Human-Centered Framework for Transparency in Fairness-Aware Recommender Systems." Proceedings of the 16th ACM Conference on Recommender Systems.
- "Social Media Lab." 2023. Fact Check Assistant_Github.
<https://github.com/smlabto/factcheckassistant>.
- Sodré, Muniz. 2019. "O facto falso: Do factóide às fake news." *As fake news e a nova ordem (des) informativa na era da pós-verdade*: 87-100.
- Soni, Jimmy, and Rob Goodman. 2017. *A mind at play: how Claude Shannon invented the information age*. New York: Simon and Schuster.
- Søraa, Roger Andre. 2017. "Mechanical genders: how do humans gender robots?" *Gender, Technology and Development* 21 (1-2): 99-115.
- Sørensen, Kristine, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, and Helmut Brand. 2012. "Health literacy and public health: a systematic review and integration of definitions and models." *BMC public health* 12 (1): 1-13.
- Sousa, Jorge Pedro. 2005. *Elementos de jornalismo impresso*. Porto: Letras Contemporaneas.
- Souza, Bruno Mello, and Simone Piletti Viscarra. 2018. "Democracia e polarização eleitoral no Brasil: as opiniões dos Eleitores De PT e do PSDB (2002-2014)." *Missões: Revista de Ciências Humanas e Sociais* 4 (3): 29-44.
- Spiekermann, Sarah. 2015. *Ethical IT innovation: A value-based system design approach*. Boca Raton: CRC Press.
- Spyridakis, JH, and JR Fisher. 1992. "Usability testing in technical communication: The application of true experimental designs." *Technical Communication*: 469-472.
- Srba, Ivan, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. "Monant medical misinformation dataset: Mapping articles to fact-checked claims." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.
- Strauss, Anselm, and Juliet M Corbin. 1997. *Grounded theory in practice*. London: Sage.
- Stray, Jonathan. 2021. "Making artificial intelligence work for investigative journalism." *Algorithms, Automation, and News*: 97-118.
- Sutrop, Margit, and Katrin Laas-Mikko. 2012. "From identity verification to behavior prediction: Ethical implications of second generation biometrics." *Review of policy research* 29 (1): 21-36.
- Suwajanakorn, Supasorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. "Synthesizing obama: learning lip sync from audio." *ACM Transactions on Graphics (ToG)* 36 (4): 1-13.
- Sylvia Chou, Wen-Ying, Anna Gaysynsky, and Joseph N Cappella. 2020. "Where we go from here: health misinformation on social media." 110 (S3): S273-S275.
- "Synthesia." 2023. Create professional videos without mics, cameras, or actors.
<https://www.synthesia.io>

- Taeihagh, Araz. 2021. "Governance of artificial intelligence." *Policy and society* 40 (2): 137-157.
- Tandoc Jr, Edson C, Zheng Wei Lim, and Richard Ling. 2018. "Defining "fake news" A typology of scholarly definitions." *Digital journalism* 6 (2): 137-153.
- Tariq, Shahroz, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. 2018. "Detecting both machine and human created fake face images in the wild." Proceedings of the 2nd international workshop on multimedia privacy and security.
- Tashtoush, Yahya, Balqis Alrababah, Omar Darwish, Majdi Maabreh, and Nasser Alsaedi. 2022. "A deep learning framework for detection of COVID-19 fake news on social media platforms." *Data* 7 (5): 65.
- Tenbrunsel, Ann E, and David M Messick. 2004. "Ethical fading: The role of self-deception in unethical behavior." *Social justice research* 17: 223-236.
- Thorson, Emily. 2016. "Belief echoes: The persistent effects of corrected misinformation." *Political Communication* 33 (3): 460-480.
- "TITAN." 2023. Project: Engineering_Ingegneria. <https://www.titanthinking.eu>.
- Tong, Jingrong, and Landong Zuo. 2021. "The inapplicability of objectivity: Understanding the work of data journalism." *Journalism Practice* 15 (2): 153-169.
- Träsel, Marcelo Ruschel. 2014. "Entrevistando planilhas: estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no Brasil. Tese de Doutorado." PhD Thesis., Pontifícia Iniversidade Católica do Rio Grande do Sul.
- Trevino, Linda Klebe. 1986. "Ethical decision making in organizations: A person-situation interactionist model." *Academy of management Review*. 11 (3): 601-617.
- Trokhymovych, Mykola, and Diego Saez-Trumper. 2021. "Wikicheck: An end-to-end open source automatic fact-checking api based on wikipedia." Proceedings of the 30th ACM International Conference on Information & Knowledge Management.
- Tukey, John W. 1977. *Exploratory data analysis*. Vol. 1. London: Pearson.
- Turek, Matt. 2019. "Media Forensics (MediFor)." *Defense Advanced Research Projects MediFor*. 17.
- Umbrello, Steven. 2021. "Towards a Value Sensitive Design Framework for Attaining Meaningful Human Control over Autonomous Weapons Systems."
- Van de Poel, Ibo. 2020. "Embedding values in artificial intelligence (AI) systems." *Minds and Machines*. 30 (3): 385-409.
- Van der Haak, Bregtje, Michael Parks, and Manuel Castells. 2012. "The future of journalism: Networked journalism." *International journal of communication* 6: 16.
- van der Stappen, Esther, and Marlies van Steenbergen. 2020. "The Ethical Matrix in Digital Innovation Projects in Higher Education." BLED 2020 Proceedings.
- van Steenbergen, Marlies, and Irene van der Spoel. 2021. "Online Proctoring: Adding Human Values to the Equation." LED 2021 Proceedings.
- van Steenbergen, Marlies; Timmers, Cathelijn 2022. "Applying Ethics in Digital Innovation." Utrecht University.
- Vedula, Nikhita, and Srinivasan Parthasarathy. 2021. "Face-keg: Fact checking explained using knowledge graphs." Proceedings of the 14th ACM International Conference on Web Search and Data Mining.
- Villar Rodriguez, Guillermo, Monica Maria Souto Rico, Alejandro Martín, Guillermo Villar Rodríguez, and Mónica María Souto Rico. 2022. "Virality, only the tip of the iceberg: ways of spread and interaction around COVID-19 misinformation in Twitter." 35 (2): 239-256.

- Vlachos, Andreas, and Sebastian Riedel. 2014. "Fact checking: Task definition and dataset construction." Proceedings of the ACL 2014 workshop on language technologies and computational social science.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359 (6380): 1146-1151.
- Vraga, Emily K, and Leticia Bode. 2020. "Correction as a solution for health misinformation on social media." *American Journal of Public Health (AJPH)* 110 (3): 278-280.
- "VSD Lab." 2020. Accessed April, 1st. <https://vsdesign.org>.
- Wang, Junyang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, and Jihua Zhu. 2023. "Evaluation and Analysis of Hallucination in Large Vision-Language Models." *arXiv preprint arXiv:2308.15126*.
- Wang, Xin, Fan Chao, and Guang Yu. 2021. "Evaluating rumor debunking effectiveness during the COVID-19 pandemic crisis: utilizing user stance in comments on Sina Weibo." *Frontiers in public health* 9: 770111.
- Wang, Xuya, and Feng Zhu. 2022. "The application of artificial intelligence in AI news anchor." International Conference on Big Data Analytics for Cyber-Physical System in Smart City.
- Ward, Stephen JA. 2019. "Journalism ethics." In *The handbook of journalism studies.*, 307-323. London: Routledge.
- Wardle, Claire, and Hossein Derakhshan. 2018. "Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information." *Journalism, 'fake news' & disinformation*: 43-54.
- Wasike, Ben. 2022. "Memes, memes, everywhere, nor any meme to trust: Examining the credibility and persuasiveness of COVID-19-related memes." *Journal of Computer-Mediated Communication* 27 (2): 1-30.
- Wei, Dengjie. "Construction of a Digital Color Grading Laboratory Based on DaVinci Resolve." *The Frontiers of Society, Science and Technology*. 5 (14): 84-89.
- Weick, KE. 2005. *Making sense of blurred images: Mindful organizing in Mission STS-107*. WH Starbuck, M. Farjoun, eds. *Organization at the Limit: Lessons from the Columbia Disaster*. Oxford: Blackwell Publishing.
- Werbos, Paul J. 2006. "Backwards differentiation in AD and neural nets: Past links and new opportunities." *Automatic differentiation: Applications, theory, and implementations.*: 15-34.
- Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers." *Science* 131 (3410): 1355-1358.
- Woelfer, Jill Palzkill, Amy Iverson, David G Hendry, Batya Friedman, and Brian T Gill. 2011. "Improving the safety of homeless young people with mobile phones: Values, form and function." Proceedings of the SIGCHI conference on human factors in computing systems.
- Woloszyn, Vinicius, Eduardo G Cortes, Rafael Amantea, Vera Schmitt, Dante AC Barone, and Sebastian Möller. 2021. "Towards a novel benchmark for automatic generation of claimreview markup." Proceedings of the 13th ACM Web Science Conference 2021.
- Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. "Bloomberggpt: A large language model for finance." *arXiv preprint arXiv:2303.17564*.

- Wu, Yangjian, and Gang Hu. 2023. "Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings." Proceedings of the Eighth Conference on Machine Translation.
- Wührl, Amelie, and Roman Klinger. 2021. "Claim detection in biomedical Twitter posts." *arXiv preprint arXiv:2104.11639*.
- Xavier, Valêncio. 1998. *O mez da gripe e outros livros*. São Paulo: Companhia das Letras.
- Yang, Kai-Cheng, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. "Arming the public with artificial intelligence to counter social bots." *Human Behavior and Emerging Technologies*. 1 (1): 48-61.
- Yao, Jiarui, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. "Factuality assessment as modal dependency parsing." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.
- Yao, Yuanshun , Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. 2017. "Automated Crowdturfing Attacks and Defenses in Online Review Systems." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.
- Yoo, Daisy, Alina Huldtgren, Jill Palzkill Woelfer, David G Hendry, and Batya Friedman. 2013. "A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts." Proceedings of the SIGCHI conference on human factors in computing systems.
- Yuan, Qianli, and Mila Gasco-Hernandez. 2021. "Open innovation in the public sector: creating public value through civic hackathons." *Public Management Review*. 23 (4): 523-544.
- Zarocostas, John. 2020. "How to fight an infodemic." *The lancet*. 395 (10225): 676.
- Zhu, Biru, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. "Knowledge enhanced fact checking and verification." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 29: 3132-3143.
- Zuboff, Shoshana. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. London: Profile books.
- Zuo, Chaoyuan, Ayla Karakas, and Ritwik Banerjee. 2018. "A hybrid recognition system for check-worthy claims using heuristics and supervised learning." CEUR workshop proceedings.

ANNEXURES AND APPENDICES

Annex A – Academic Path 2020-2024

SCIENTIFIC EXPEDITION

2024/03/11 – 2024/07/15 – Lund University

Description:

Visiting Fellow to update my knowledge of the concerns and goals of Automatic Journalism, fact-checking, newscasts, and research centers working on media innovation, including developing regional and international networks. I would highlight the partnership with the Department of Sociology, the areas of Design Sciences at Lund University of Technology (LTH), and the Centre for Innovation Research (CIRCLE). Lunds universitet Sociologiska institutionen, Suécia

2022/08/22 - 2022/08/27 - Utrecht University

Description:

Programa Erasmus+ (short mob PhD) - Summer School
Utrecht Summer School in the course Applying Ethics in Digital Innovation.
Theologische Universiteit Utrecht, Netherlands

PUBLICATIONS

Conference

- 1 Maia, H. C. "Disinformation and health: A systematic review study on health automatic fact-checking during the pandemic". Trabalho apresentado em *17th International AAAI Conference on Web and Social Media, Limassol, 2023*.
[https://workshop-proceedings.icwsm.org/pdf/2023_42.pdf]
- 2 Maia, H. C. "Tracing Health Communication: a systematic review about apps under covid 19 Pandemic". Trabalho apresentado em *9th European Communication Conference, Aarhus, 2022*.
[https://conferences.au.dk/fileadmin/conferences/2022/ECREA/FINAL_Electronic_booklet_ECREA.pdf]
- 3 Maia, H. C.; Essenfelder, Renato; Nodari, Sandra. Autor correspondente: Maia, H. C.. "An approach for expected roles and commonsense through journalists and AI developers: a Delphi study in Brazil". Trabalho apresentado em *The International Association for Media and Communication Research - IAMCR, Beijing, 2022*. [<https://iamcr.org/beijing2022/abstract-books>]
- 4 Foreiro, Jorge; Maia, Haline Costa; Casarin, Jordana. C.. "Extended researchers: Towards a meta social human beings". Trabalho apresentado em *LINK 2021 3rd International Conference on Practice-Oriented Research in Art & Design, Auckland, 2021*.
[<https://doi.org/10.24135/link2021.v2i1.174>]

Journal

- 1 Maia, H. C.. 2024. "Automated Journalism and the ethnographic approach applied within the Brazilian Sports Newsroom". *Journal of Advanced Media and Communication*:
<https://www.inderscience.com/jhome.php?jcode=ijamc> | Entering Publication Schedule
- 2 Maia, H. C., Ariel, C. P. and Nunes, S. 2024. "Adding Human Values on the Deepfake: Co-Designing Fact-Checking Solutions to Combat Misinformation". *AI and Ethics*. Springer Nature:
<https://link.springer.com/journal/43681> | Approved for publication – Under revision

Book Chapter

- 1 Maia, H. C.. "Jornalismo na Era de Novos Ecosistemas Tecnológicos no Grupo Globo". #7COBCIBER, Porto, 2022. Published: [<https://doi.org/10.21747/978-989-8970-77-0/cib>]
<https://ler.letras.up.pt/site/default.aspx?qry=id024id1937&sum=sim&n0=Edi%C3%A7%C3%B5es%20do%20CITCEM&n1=Ciberjornalismo>]

2 Nodari, Sandra; Maia, Haline; Vernizi, Eduardo. "Press offices reinvention after Operation Car Wash (Brazil): critical and practical perspectives". *Estudos Mediáticos: memória VI*. Portugal, 2023. Published [<http://hdl.handle.net/10362/164627>]

3 Coca, Adriana; Essenfelder, Renato; Maia, Haline Costa. Autor correspondente: Coca, Adriana. "Mulheres de Abril: a Revolução dos Cravos traduzida em um docudrama com olhar feminino.". In *Jornalismo e Estudos Mediáticos: memória V*. Portugal, 2022. Published.

CONFERENCES

Oral Presentation

1. Maia, H. C.. "Evaluating the Ethical Implications of Deepfake Use in Journalism using Value Sensitive Design and Automated Fact-Checking". Trabalho apresentado em *C+J Data Journalism Conference, Zurich, 2023*.
2. Maia, H. C.. "Disinformation and health: A systematic review study on health automatic fact-checking during the pandemic". Trabalho apresentado em *17th International AAAI Conference on Web and Social Media, Limassol, 2023*.
3. Maia, H. C.. "Tracing Health Communication: a systematic review about apps under covid 19 Pandemic". Trabalho apresentado em *9th European Communication Conference, Aarhus, 2022*.
4. Maia, H. C.; "An approach for expected roles and commonsense through journalists and AI developers: a Delphi study in Brazil". Trabalho apresentado em *The International Association for Media and Communication Research - IAMCR, Beijing, 2022*.
5. Maia, H. C.. "Jornalismo na Era de Novos Ecossistemas Tecnológicos no Grupo Globo". Trabalho apresentado em *LUSOCOM - Federação Lusófona de Ciências da Comunicação, Porto, 2022*.
6. Maia, H. C.; "Extended researchers: Towards a meta social human beings". Trabalho apresentado em *LINK 2021 3rd International Conference on Practice-Oriented Research in Art & Design, Auckland, 2021*.

EVENT ORGANIZATION

2023/06/15 - 2023/06/16 - 5th FEUP Doctoral Congress Engineering (DCE23)

Description:

I was a member of the Symposium on Informatics organizing committee, held in the scope of the 5th FEUP Doctoral Congress Engineering (DCE23). DCE23 is an opportunity for doctoral students, of the several areas of engineering, to discuss their on-going research with peers, professors and industry/companies and also for potential industry partners to contact a broad audience of engineers which will soon be looking to move on to the industry.

2021/12/01 - 2021/12/02 - LINK 2021 3rd International Conference on Practice-led research in Art and Design

Description:

As an onboard member of the conference, I was responsible for the review of submissions and the invitation of participants. This involved meticulously reviewing submitted papers, abstracts, and other materials to ensure they met the conference's standards and thematic focus. Based on the accepted submissions, I extended formal invitations to authors, providing clear and professional communication, including details on presentation formats, schedules, and logistical arrangements.

2021/06/15 - 2021/06/16 – In Bloom

Description:

IN BLOOM is a workshop organized by first years' PhD students from the Doctoral Program in Digital Media of the University of Porto. The initiative proposes a conversation about topics along the Digital Media field related to new technologies, media, design, computing, and other areas. Besides, the workshop is an opportunity to share knowledge in research and exchange experiences with the scientific community around the world. This year, the workshop sought to explore new perspectives and ask questions that can lead to better solutions.

GUEST LECTURER

2023/10/03 - 2023/10/15 - Digital Media in Health Communication and Literacy

Description:

This course gave participants a comprehensive understanding of digital literacy, equipping them with the skills and knowledge needed to navigate and contribute to our increasingly digital world. My webinar provides background information and prepares you with specific knowledge that will be extremely helpful for your project development and your Hackathon Experience's success.

Link: <https://www.hackathonpractices.com/5th-edition-digital-communication-and-literacy>

2020/11/27 - 2021/02/19 - EUGLOH - European University Alliance for Global Health

Description:

I was a mentor in the Digital Media in Health Communication and Literacy, a fully online program that aims at bringing together EUGLOH students with different perspectives and skills to create collaborative digital media-based projects that communicate complex concepts in scientific research. During the intensive 11 week Hackaton, participants are immersed in a collaborative setting where they are guided through the process of exploring different methodologies (e.g. design thinking, lean startup, Agile) with a focus on creating health communication and literacy.

ASSISTANT PROFESSOR

2024/09/01 - Present

IPAM · Instituto Português de Administração e Marketing

Full-time -, Portugal · On-site

Description:

As an Assistant Professor at IPAM, Portugal, I teach International Economics and Business, leveraging IPAM's modern, student-centered pedagogical approach. My role involves developing and delivering dynamic, practice-oriented courses that prepare students for the complexities of the global business landscape. Engaging in both teaching and collaborative research, I contribute to fostering a forward-thinking academic environment that bridges theory and real-world application.

○

AFFILIATE RESEARCHER

2024/09/25 - Present

CIRCLE, Lund University - CIRCLE · Remote - Lund, Skåne County, Sweden

○

CIRCLE is the Centre for Innovation Research at Lund University. It functions as incubator, coordinator, and implementer of innovation research across faculties. Our aim is to understand and explain how innovation can contribute to a good society and tackle societal challenges such as economic crises, climate change or the increased globalization of economic activities.

The research is organized into nine research themes, hosting different research projects, and most researchers at CIRCLE are involved in several themes. These are the activities realized:

- Published & working papers promoted through CIRCLE communication channels;
- Participate in projects at CIRCLE and apply to use data bases;
- Support communication activities (e.g. retweet relevant work, etc.);
- CIRCLE regular open research activities (e.g. research talks, seminars);
- Report all relevant research and outreach activities.

Annex B

INESCTEC Data Storage

Field	Value
Source	Data from fact-checking initiatives
Contributor	Priscila Ariel
DOI	https://doi.org/10.25747/TKJR-EV33
Data Collection Method	Data generated using API and prototypes
Format	mp4
Language	PT and EN
Temporal Coverage	From 2022 to 2024
Type	Outputs of fact-checking videos using various LLM applications
Type of Instrument	LLM applications. Globo prototype, Akool, D-ID, GPT, Claude

File Name	Thesis Section	Link Name
AIDA	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/66041606bfea1d00f25a3c01
GAB	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/6604158aeb56f0010a8a3ade
Synthesia	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/660415bdeb56f001128a38d3
D-ID	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/660415d901fc9c00ea3171c8
GAB creation proces	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/66041589bfea1d00f55a3c48
AIDA creation proc	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/66041607bfea1d00f65a3e58
D-ID Creation Proces	Workshop 1 and 2	https://osf.io/b853a/files/osfstorage/660415dabfea1d00f65a3d2e
AIDA - portuguese	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27cb821e8425e9e24024a
Bigodão	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27ced76fd535a0116757
Dina Aguiar	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27ccd037c2aeb4d240927
Melinda	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27cdcf1943aa42a80d5f5
Polígrafo	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27d0ad12013490180d15d
Sônia Rebelo	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27ca63bee8747c380d9f0
Text and Audio only	Workshop 3	https://osf.io/dacqm/files/osfstorage/66a27cfabc94ba7c179101d2
3 Minutes Thesis	Vídeo Abstract	https://osf.io/zmxgh/files/osfstorage/66a3b4850862cfb5e7c9a626

Appendix A

The systematic Review paper details:

Categories	Quantity	Autors	Techniques
checking	25	(Liu et al., 2022)	- Detect falser news in the Chinese language using four criteria: source, author, message, and spelling. Reliable human feedback can help detect false news with high accuracy.
		(Tashtoush et al., 2022)	- Long Short-Term Memory (LSTM), Bi-directional LSTM, Convolutional Neural Network (CNN), and a hybrid of CNN and LSTM networks, to automatically classify and identify fake news content related to the COVID-19 pandemic.
		(Brand et al., 2022)	- E-BART provide a veracity prediction for a claim, and jointly generate a human-readable explanation for this decision.
		(Aloshban, 2020)	- Automatic Fake News Classification Through Self-Attention (ACT), which exploits different supportive articles to a claim which mimics manual fact-checking processes.
		(Nakov et al., 2021)	
		(Botnevik et al., 2020)	- BRENDA, a browser extension which uses a tested deep neural network architecture to automatically identify fact check worthy claims and classifies as well as presents the result along with evidence to the user.
		(Elhadad et al., 2021)	- Automatically annotated, bilingual (Arabic/English) COVID-19 Twitter dataset (COVID-19-FAKES). This dataset help in studies related to sentiment analysis, the analysis of the propagation of misleading information related to this outbreak, the analysis of users' behavior during the crisis, the detection of botnets, the analysis of the performance of different classification algorithms with various feature extraction techniques that are used in text mining.
		(Pathak et al., 2021)	- Dataset that provide error analysis that illustrates the challenges involved in the automated fact-checking task and identifies factors that may improve this process in future work. Also describe the implementation of a web app that extracts important entities and actions from a given article and searches the web to gather evidence from credible sources.
		(Hammouchi & Ghogho, 2022)	- Framework for detecting fake news that uses external evidence to verify the veracity of online news in a multilingual setting with Covid-19 related news.
		(Sarrouti et al., 2021)	.- HEALTHVER, a new dataset for evidencebased fact-checking of health-related claims that allows to study the validity of real-world claims by evaluating their truthfulness against scientific articles.
		(Bonet-Jover et al., 2021)	- A new dataset was created and annotated with a new fine-grained annotation scheme (FNDeepML) that considers the different elements of the news document and their veracity about health.
		(Vedula & Parthasarathy, 2021)	- FACE-KEG, to automatically perform explainable fact checking. Given an input fact or claim, our proposed model constructs a relevant knowledge graph for it from a large-scale structured knowledge base.
(Jiang et al., 2020)	- Factoring fact-checks for automatically extracting structured information from fact-checking articles.		
(Yao et al., 2021)	- Frame factuality assessment as a modal dependency parsing task that identifies the events and their sources,		

		formally known as conceivers, and then determine the level of certainty that the sources are asserting with respect to the events (NLP applications such as fact checking and rumor detection).
	(Charles et al., 2022)	- Newsclassification experiment using the LSTM model, and we achieved great results, with a precision of 97\% on classifying fake news and 92\% on classifying real news.
	(Saquete et al., 2020)	
	(Karagiannis et al., 2020)	- Data set containing over 110,000 ranked mistakes with a precision of 85% in the top 1% and a precision of over 60% in the top 25%; AKB can be used to find mistakes on the entire Web.
	(Shaar et al. 2021)	- The proceedings contain 21 papers: identifying automatically generated headlines using transformers; improving hate speech type and target detection with hateful metaphor features; improving cross-domain hate speech detection by reducing the false positive rate; understanding the impact of evidence-aware sentence selection for fact checking; leveraging community and author context to explain the performance and bias of text-based deception detection models; never guess what I heard... rumor detection in Finnish news: a dataset and a baseline; extractive and abstractive explanations for fact-checking and evaluation of news; generalizability of topic models in cross-corpora abusive language detection; AraStance: a multi-country and multi-domain dataset of Arabic stance detection for fact checking; an empirical assessment of the qualitative aspects of misinformation in health news; and findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection.
	(Springstein et al., 2021)	- Web-based demo application that automatically quantifies the cross-modal relations of entities (persons, locations, and events) in image and text. The applications are manifold.
	(Smeros et al., 2021)	- SciClops involves three main steps to process scientific claims found in online news articles and social media postings: extraction, clustering, and contextualization.
	(Simpson & Yang, 2022)	.- Fact-checking in an automatic manner, which autonomously discerns the veracity of a given statement quickly: supervised machine learning approaches for this intended purpose and how they interact with numerical features extracted from a dataset of journalistic writing focusing on the efficacy and efficiency thereof.
	(Arnold, 2020)	
	(Woloszyn et al., 2021)	- Artificial Intelligence for generating ClaimReview automatically.
	(Azevedo et al., 2022)	- Recall and precision in the retrieval of documents and sentences that can be provided to justify the veracity of a given claim: document Retrieval task using coreference resolution to improve the Sentence Retrieval task.
	(Mykola & Saez-Trumper, 2021)	- Propose a data filtering method that improves the model's performance and generalization. Then, the WikiCheck API that automatically performs a facts validation process based on the Wikipedia knowledge base.
monitoring	18	(Moravec et al., 2020) - Investigate two different approaches to implementing a fake news flag-one designed to have its primary effect when processed by automatic cognition (System 1) and

	the other designed to have its primary effect when processed by deliberate cognition (System 2).
(Dai et al., 2020)	- Propose elucidating fact-checking predictions using counterfactual explanations to help people understand why a specific piece of news was identified as fake. Experiments are conducted on the FEVER dataset for both system and human evaluations.
(Nakov et al., 2021)	
(Casillas et al., 2022)	- Scores computed within the attention layer to show which evidence spans are more relevant to classify a claim as supported or refuted. Our classification models achieve results compared to the state-of-the-art models in terms of classification of relevance evaluation and claim verification accuracy on the FEVER dataset.
(Savage, 2021)	- Artificial intelligence provides automatic fact-checking and fake news detection, but with limits.
(Nabožny et al., 2021)	- Fact-checking performance of expert annotators on capturing misinformation by the factor of 2.2 on average. and allows medical experts to fact-check and identify over two times more non-credible medical statements.
(Moffitt, 2021)	- This model is based on the Bidirectional Encoder Representations from Transformers (BERT) model developed by Google researchers, that expedites analysis by automating a process that is currently done manually (identifying tweets that promote COVID-19 conspiracy theories).
(Neumann et al., 2022)	- This framework should help researchers, policymakers, and practitioners reason about potential harms or risks associated with these algorithms and provide conceptual guidance for the design of algorithmic fairness audits in this domain.
(Rakotoson et al., 2022)	- Final system on an end-to-end Extractive Question Answering (EQA) combined with a three outputs classification model to perform in-depth semantic understanding of a question to illustrate the aggregation of multiple responses.
(Ding et al., 2022)	- MetaDetector pushes the feature extractor and event discriminator to eliminate event-specific features and preserve required meta knowledge by adversarial training. Furthermore, the pseudo-event discriminator is utilized to evaluate the importance of news records in historical events to obtain partial knowledge that are discriminative for detecting fake news.
(Srba et al. 2022)	- Feature-rich dataset of approx. 317k medical news articles/blogs and 3.5k fact-checked claims. It also contains 573 manually and more than 51k automatically labelled mappings between claims and articles.
(Shaar et al. 2021)	- The proceedings contain 21 papers: identifying automatically generated headlines using transformers; improving hate speech type and target detection with hateful metaphor features; improving cross-domain hate speech detection by reducing the false positive rate; understanding the impact of evidence-aware sentence selection for fact checking; leveraging community and author context to explain the performance and bias of text-based deception detection models; never guess what I heard... rumor detection in Finnish news: a dataset and a baseline; extractive and abstractive explanations for fact-checking and evaluation of news; generalizability of topic models in cross-corpora abusive language detection; AraStance: a multi-country and multi-domain dataset of Arabic stance detection for fact checking; an empirical assessment of the qualitative

			aspects of misinformation in health news; and findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection.
		(Chen & Freire, 2020)	- Through this interface, fact-checkers can visualize domain interaction networks, query domain fakeness score, and tag incorrectly predicted results. Our proactive discovery system will expedite fact-checking process and can be a powerful weapon in the toolbox to combat misinformation.
		(Reis & Benevenuto, 2021)	- Automatic approaches based on supervised machine learning as a support tool to help fact-checkers identify misinformation shared through images on WhatsApp.
		(Arnold, 2020)	
		(Lanius et al., 2021)	- The COVID-19 infodemic is driven partially by Twitter bots. Flagging could be built into future automated fact-checking systems and other misinformation abatement strategies of the social network analysis and mining community.
		(Lahoti & Prasad, 2022)	- Two-way approach has been taken to classify Twitter messages (Tweets) as rumor or non-rumor related. The first approach is text-based analysis, while the other is media-based analysis.
		(Villar-Rodríguez et al., 2022)	- Work combines the use of artificial intelligence and journalism through fact-checked false claims to provide an in-depth study of the number of retweets, likes, replies, quotes and repeated texts in posts stating or contradicting misinformation in Twitter.
publishing	6	(Nakov et al., 2021)	
		(Vedula & Parthasarathy, 2021)	- FACE-KEG, to automatically perform explainable fact checking. Given an input fact or claim, our proposed model constructs a relevant knowledge graph for it from a large-scale structured knowledge base.
		(Gruzd and Mai, 2020)	- Case study examines how a conspiracy theory propagated on Twitter and whether the hashtag virality was aided by the use of automation or coordination among Twitter users. While the spread of misinformation can be potentially mitigated by fact-checking and directing people to credible sources of information from public health agencies, false and misleading claims that are driven by politics and supported by strong convictions and not science are much harder to root out.
		(Simpson & Yang, 2022)	- Fact-checking in an automatic manner, which autonomously discerns the veracity of a given statement quickly: supervised machine learning approaches for this intended purpose and how they interact with numerical features extracted from a dataset of journalistic writing focusing on the efficacy and efficiency thereof. ection approaches can be used in practice, highlighting challenges and opportunities.
		(Babakar & Moy, 2021)	
		(Lanius et al., 2021)	- The COVID-19 infodemic is driven partially by Twitter bots. Flagging could be built into future automated fact-checking systems and other misinformation abatement strategies of the social network analysis and mining community.
spotting	20	(Pogorelov et al., 2021)	- A dataset that deals with the idea that the COVID-19 pandemic is causally connected to the 5G wireless network. They provide the human-annotated dataset along with an additional large-scale automatically (by using the human-annotated dataset as the training set) labelled dataset consist of more than 100,000 tweets.

(Ng & Carley, 2021)	- The COVID-19 infodemic is driven partially by Twitter bots. Flagging could be built into future automated fact-checking systems and other misinformation abatement strategies of the social network analysis and mining community.
(Sharma & Garg, 2021)	- Explore stories reported by fact-checking groups PolitiFact, Poynter and Snopes from January to June 2020. Results show story type classification performs best when trained on the same medium, with contextualised BERT vector representations outperforming a Bag-Of-Words classifier.
(Harrag and Djahli, 2021)	- Various supervised techniques of the ML model and advanced deep learning techniques are implemented for intense research. The result of these algorithms helps in differentiating false content from reliable news and improved the accuracy achieved using existing techniques.
(Alhindi et al., 2021)	- The main goal is to develop a model capable of automatically detecting truth given an Arabic news or claim. In particular, a deep neural network approach that can classify fake and real news claims by exploiting 'Convolutional Neuron Networks'.
(Nakov et al., 2021)	
(Cheema et al, 2020)	- Fusion of syntactic features and deep transformer Bidirectional Encoder Representations from Transformers (BERT) embeddings, to classify check-worthiness of a tweet, i.e. whether it includes a claim or not. Claim retrieval, we explore the pre-trained embeddings from a Siamese network transformer model (sentence-transformers) specifically trained for semantic textual similarity, and perform KD-search to retrieve verified claims with respect to a query tweet.
(Wüthrl & Klinger, 2021)	- Baseline classification experiments with embedding-based classifiers and BERT-based transfer learning demonstrate that the detection is challenging, however, shows acceptable performance for the identification of explicit expressions of claims.
(Mansour et al., 2022)	- BERT variants as point-wise rerankers. Additionally, we study the impact of using different fields of the verified claim during training and inference phases.
(Wang et al, 2021)	- Formulating effective strategies to debunk rumors and curb their spread on social media. Then, based on stance responses, a debunking effectiveness index (DEI) was developed to measure the effectiveness of different debunking methods. Findings provide relevant insights into ways to debunk rumors effectively, support crisis management of false information, and take necessary actions in response to rumors amid public health emergencies.
(Sarrouti et al., 2021)	- HEALTHVER, a new dataset for evidencebased fact-checking of health-related claims that allows to study the validity of real-world claims by evaluating their truthfulness against scientific articles.
(Kotony & Toni, 2020)	- PUBHEALTH of 11.8K claims accompanied by journalist crafted, gold standard explanations (i.e., judgments) to support the fact-check labels for claims. We explore two tasks: veracity prediction and explanation generation. We also define and evaluate, with humans and computationally, three coherence properties of explanation quality.
(Sawczyn et al., 2021)	- Leveraging recent advancements in natural language processing, acquired relevant data and developed

		classifiers of evidence relevance with respect to claims in Polish.
	(Martín et al., 2022)	- FacTeR-Check implements three different modules relying on the XLM-RoBERTa Transformer architecture to evaluate semantic similarity, to calculate natural language inference and to build search queries through automatic keywords extraction and Named-Entity Recognition.
	(Saquete et al., 2020)	
	(Carpuat & Meza-Ruiz, I.(2022)	- PubHealthTab: a public health table-based dataset for evidence-based fact checking.
	(Zhu et al., 2021)	- A knowledge enhanced fact checking system, which leverages the Wikidata5M knowledge graph and Wikipedia documents to incorporate external knowledge into the claim to be checked for more robust and accurate fact checking. First, we devise a contextualized knowledge graph selection method to identify the most
	(Shaar et al, 2020)	- Models based on BERT, LSTMs, and CNNs:datasets from the lab as well as the evaluation scripts, which should enable further research in the important tasks of check-worthiness estimation and detecting previously fact-checked claims.
	(Nikolov et al. 2020)	- Manual fact-checking is time-consuming and automatic fact-checking is resource-intense, which means that we need to pre-filter the input social media posts and to throw out those that do not appear to be check-worthy. With this in mind, here we propose a model for detecting check-worthy tweets about COVID-19, which combines deep contextualized text representations with modeling the social context of the tweet.
	(Arnold, 2020)	

Appendix C

Conference Program - Workshops

Thursday 22 June 2023	
Tamedia HQ, Werdstrasse 21, Zurich	
Workshop room 1	Workshop room 2
12:30	Registration desk opens a 12:30 at Tamedia (until 16:30)
13:00 - 13:20	Beyond the hype: an introduction to NLP and language models
13:20 - 13:40	Envisioning the Future of AI in the News Media
13:40 - 14:00	Adriano Belisario and Ajai Radhakrishnan
14:00 - 14:20	Nicholas Diakopoulos, Natali Helberger and Kimon Kieslich
14:20 - 14:40	Publishing your notebooks with the Quarto framework
14:40 - 15:00	Jonathan Soma and Leon Yin
15:00 - 15:20	What policies do you need to build a better future for AI and Journalism?
15:20 - 15:40	Who's Behind A Website?
15:40 - 16:00	Noémie Krack, Anna Schjøtt Hansen and Aleksandra Kuczerawy
16:00 - 16:20	Priyanjana Bengani and Jon Keegan
16:20 - 16:40	
16:40 - 17:00	Evaluating the Ethical Implications of Deepfake Use in Journalism using Value Sensitive Design and Automated Fact-Checking
17:00 - 17:20	Finding and using undocumented APIs
	Leon Yin and Piotr Sapiezynski
	Haline Maia

Appendix D

Contributed Workshop - Camera-ready

Title: Evaluating the Ethical Implications of Deepfake Use in Journalism Using Value Sensitive Design and Fact-Checking

Workshop Abstract (250 words):

This workshop aims to use the Value Sensitive Design (VSD) methodology to evaluate the ethical implications of deepfake use in journalism. The participants will learn about the VSD framework and how it can be used to assess the moral values of technology, in this case, deepfakes. This methodology ensures that technology aligns with human values, norms, and requirements. To obtain these analyses and data, the workshop will examine two existing cases of deepfake use in journalism, one by a legitimate news outlet and another by fake news. In the context of deepfake use in journalism, some relevant values and norms include truth, credibility, transparency, privacy, and consent. After identifying how deepfakes affect different values, requirements, and norms in journalism, we will discuss the possibility of aligning synthetic media technologies for automated fact-checking dissemination. One way to align deepfakes with values such as truth and credibility is to use the technology for fact-checking purposes. For example, a fact-checking outlet could use digital doppelgangers to create real-time footage to detect instances of manipulation or disinformation. In conclusion, using deepfakes in journalism raises important ethical and social questions. Furthermore, the workshop participants will understand the values, norms, and requirements in these discussions from cases and discuss the possibility of aligning the use of synthetic media technologies for automated fact-checking dissemination.

Target Audience: This workshop is suitable for researchers, designers, journalists, developers, and policymakers interested in understanding and evaluating the use of fake news in journalism.

Workshop Agenda:

Introduction: The facilitator will introduce the VSD methodology and its relevance in evaluating deepfake use in journalism. The participants will also learn about the cases of deepfake and its use in journalism, which will be examined during the workshop.

Case Study 1: The facilitator will present the first case study of deepfake and virtual characters used by a legitimate news outlet. The participants will be divided into groups to analyze the ethical implications of using deepfakes in this case. The groups will also identify the stakeholders involved in this use of deepfakes.

Case Study 2: The facilitator will present the second case study of deepfake use by fake news. Again, participants will be divided into groups to analyze the ethical implications of using deepfakes in this case. The groups will also identify the stakeholders involved in this use of deepfakes.

Automated Fact-Checking: The facilitator will introduce the concept of automated fact-checking using synthetic media. Participants will be divided into groups to discuss the potential benefits and challenges of using such tools to combat disinformation and presenting values, norms, and requirements for this purpose.

Group Pitching and Discussion: The groups will reconvene to share their findings and discuss the ethical implications of deepfake use in journalism. The facilitator will guide the discussion to ensure all stakeholders and values are considered.

Conclusion: The workshop will end with a summary of the key takeaways, including the importance of using ethical frameworks such as VSD to evaluate emerging technologies and the potential of automated fact-checking tools in combatting disinformation.

Objectives:

- To understand the Value Sensitive Design methodology and its application in evaluating cases of fake use in journalism.
- To analyze two existing cases of fake use in journalism, one by a news agency and the other by fake news, and evaluate them based on values, norms, and requirements.
- To discuss the possibility of aligning the use of synthetic media technologies for automated fact-checking dissemination.
- To run simulations of these scenarios using API that create virtual characters.

Participants: Up to 12 people will be divided into groups of 3-4 people.

Facilitator:

Haline Maia:

Biographical sketch: Ph.D. Candidate at Digital Media in the Faculty of Engineering of the University of Porto, Portugal, granted with Portuguese FCT - Foundation for Science and Technology. As a Communication Specialist, she holds an MSc in Journalism from Fernando Pessoa University (Portugal) and more than ten years of experience in companies such as Globo Television, ExxonMobil, and Accenture with planning, PR, movie screenwriting, copywriting, cinema, journalism, digital marketing, and ad campaigns. Her preferred areas in academic research are journalism, NLP, artificial intelligence, communication, linguistics, technology, synthetic media, and Human-Computer Interaction (HCI).

Technical content support:

Priscila Ariel: Senior Technology and Innovation Researcher at Globo (Brazilian media company) and Ph.D. Candidate in Computer Graphics at Pontifical Catholic University of Rio de Janeiro, Brazil. She has a Master's and Bachelor's in Computer Science from the State University of Rio de Janeiro, Brazil. She has more than ten years of experience in software development. Her research interests are computer graphics, virtual reality, augmented reality, virtual production, virtual humans, and artificial intelligence.

Sérgio Nunes: Assistant Professor at the Department of Informatics Engineering at the Faculty of Engineering of the University of Porto (FEUP) and a Senior Researcher at the Centre for Information Systems and Computer Graphics at INESC TEC. His research interests are Information Retrieval and Web Information Systems, using temporal features for ranking, studying information dynamics on the web, and computational journalism.

ORCID: 0000-0002-2693-988X

Materials:

- Whiteboard and markers
- Laptops or computers
- Multimedia projection equipment
- Access to the internet
- Relevant research papers and articles
- Case studies of deep learning use in journalism

Duration: 1,5 hours

Workshop Outline:

1. Introduction (10 minutes)
 - Introduce the workshop and its objectives.
 - Provide an overview of the value-sensitive design methodology.
 - Divide groups and ask participants to take notes.

2. Explain technologies (10 minutes)
 - Introduce the concept of synthetic media, deep learning, deepfake, and virtual characters.
 - Present advantages and disadvantages of the technologies used for virtual characters.
3. Case Study 1: News Agency (5 minutes)
 - Present the case of the news agency that used deep learning technologies with a newscast avatar.
4. Case Study 2: Fake News (5 minutes)
 - Present the case of deepfake do share fake news.
5. Discussion (10 minutes)
 - Identify the stakeholders involved in these cases.
 - Lead a group discussion to identify the values, norms, and requirements that are important to consider when evaluating the use of deep learning in journalism.
 - Encourage participants to share their experiences, opinions, and perspectives.
 - Explore possible solutions and approaches to addressing the problem of fake news.
6. Aligning Synthetic Media Technologies for □Fact-Checking (30 minutes)
 - A brief explanation about automated fact-checking dissemination challenges.
 - Group Discusses how these technologies can be aligned with the values, norms, and requirements identified in the previous discussions.
 - Demonstrate using APIs such as Wonder Studio, Synthesia, and Globo Solutions to create □virtual characters simulations of these scenarios.
 - Divide participants to discuss the potential benefits and challenges of using synthetic media technologies to generate automatic and personalized content for □fact-checking dissemination.
 - Each group prepares its VSD ideas for this solution.
7. Evaluation and Conclusion (20 minutes)
 - Explain the purpose and process of pitching, such as selling an idea or presenting a summary concisely and persuasively.
 - Set the criteria for presentation: each group will have up to one minute to pitch their main points, and then we will have a brief Q&A session for clarification or elaboration.
 - End the session with a summary of the feedback, a thank-you message, and an invitation to stay in touch.

Case Studies Possibilities and Approach:

To evaluate the two cases, we will be using the following steps of the VSD methodology:

- Identify the stakeholders: In this step, we will identify the stakeholders affected using deepfakes in journalism. These stakeholders may include journalists, news outlets, audiences, the subjects of deepfakes, and society.
- Elicit stakeholder values: Once we have identified the stakeholders, we will define their possible core values.
- Instruct each group to use the value-sensitive design methodology to evaluate the case study by discussing the following questions:
 - What values are at stake in this case?
 - What norms and requirements should be met in this case?
 - How might the use of synthetic media technologies affect these values, norms, and requirements?
- Each group should take notes and write down their conclusions. The case studies possibilities are:

Case Study 1 possibilities

- In 2021, South Korean broadcaster SBS used a deepfake of its anchor to report the news. The deepfake was created using machine learning and was designed to deliver the news more naturally and engagingly. The use of the deepfake was controversial, with some praising it as a new way of delivering the news, while others saw it as a threat to the credibility of journalism.
- China's AI News Anchor: Xinhua News Agency, China's state-run news agency, introduced its first AI news anchor in November 2018. The anchor, Xin Xiaomeng, was created using deepfake technology and a virtual news anchor system. Xin Xiaomeng was designed to resemble a real-life news anchor and can read news stories with a natural-looking voice and facial expressions.
- BBC's Synthetic Newsreader: In May 2019, the BBC introduced a synthetic newsreader called "BBC's Synthetic Newsreader." The newsreader was created using deepfake technology and can deliver news in a natural-sounding voice with realistic facial expressions.
- In 2019, the Brazilian network TV Globo created a virtual news anchor named "AIDA" and "Gab" using deep learning technology. They can deliver news reports and interviews in Portuguese, and they were designed to mimic the movements and expressions of a real human anchor.

Case Study 2 possibilities

- The use of deepfakes by fake news: In 2018, a video of former US President Obama went viral on social media. The video, created using deepfake technology, showed Obama delivering a speech he had not given. The video was intended to deceive viewers and spread false information.
- Video of Nancy Pelosi manipulated to make her appear drunk: https://www.washingtonpost.com/video/politics/manipulated-video-of-house-speaker-nancy-pelosi-goes-viral-sparks-concerns/2019/05/24/7f60879e-7e7d-4c38-9d33-562c8cd2c2aa_video.html Article discussing the video: <https://www.washingtonpost.com/technology/2019/05/24/faked-pelosi-videos-slowly-spread-across-social-media-globe/?arc404=true>
- In 2020, a deepfake video featuring former UK Prime Minister Theresa May endorsing her successor, Boris Johnson, went viral on social media. A comedy channel created the video, but it raised concerns about the potential for deepfakes to be used to spread disinformation during elections.
- In 2020, a deepfake video of Facebook CEO Mark Zuckerberg was posted to Instagram. The video showed Zuckerberg talking about Facebook's power and disinformation's dangers. Still, it was created as part of an art installation to highlight the dangers of deepfakes.
- In 2020, a deepfake video of Brazilian President Jair Bolsonaro was posted on YouTube. The video showed Bolsonaro urging his supporters to engage in violence against political opponents, and it was later revealed to be a fake.
- In 2018, a deepfake video of Brazilian presidential candidate Fernando Haddad was shared on social media. The video showed Haddad making derogatory comments about the Catholic Church, and it was later revealed to be a fake.
- In 2018, a deepfake video of former President Luiz Inácio Lula da Silva was shared on social media. The video showed Lula criticizing his own Workers' Party, and it was later revealed to be a fake.
- In 2020, a deepfake video of Portuguese Prime Minister António Costa was posted on Twitter. The video showed Costa making comments about the COVID-19 pandemic that were later revealed as fake.

- In 2019, a deepfake video of Portuguese soccer star Cristiano Ronaldo was shared on social media. The video showed Ronaldo endorsing a product that he had never actually supported.
- In 2020, a deepfake video of Kim Jong Un was posted on YouTube. The video showed the North Korean leader dancing and rapping, and it was later revealed to be a fake.
- In 2020, a deepfake video of Queen Elizabeth II was posted on YouTube. The video showed the Queen making comments about the COVID-19 pandemic that were later revealed as fake.

Appendix E

Exercise:

Tamedia, a media company, plans to create a digital news verification system using deep learning technology to provide fact-checking through a news feed. They have tasked your team with evaluating the feasibility of this solution virtual interface using VSD methodologies.



Goal: We need to discuss how fact-checking technology can be aligned with the values, norms, and requirements identified in the previous discussions.

Gather in small groups and discuss the impact of deepfake and virtual characters to propose automatic fact-checking dissemination.

You have 30 minutes.

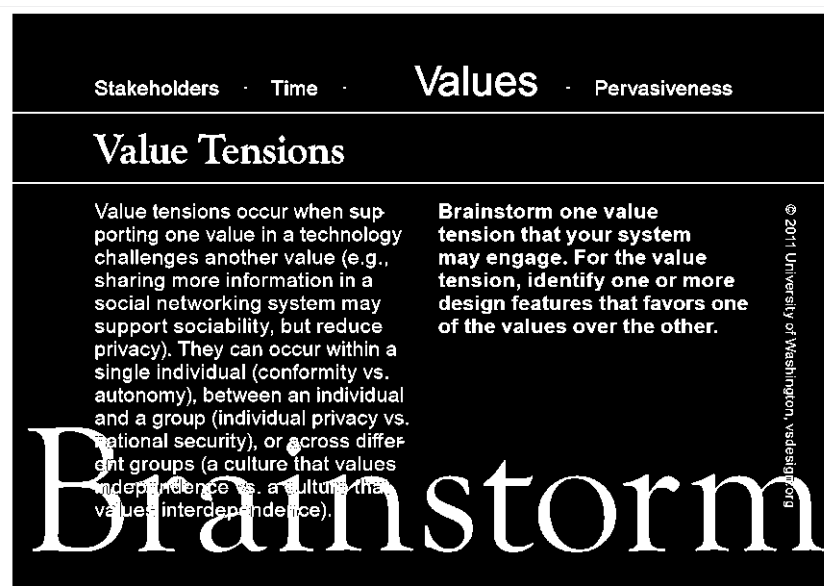
TASK 1: What are the potential harms and benefits to the identified stakeholders?

Questions to discuss:

- Identify and take notes about the videos on how to use Virtual Characters to present fact-checking.
- Share your opinions about the benefits and harms of this use.

TASK 2: Write central values and norms (at least three essential values and norms). You can take references from the value card deck provided:

TASK 3: Select one value and analyze tensions according to the envisioning cards' instructions below:



Stakeholders · Time · Values · Pervasiveness

Value Tensions

Value tensions occur when supporting one value in a technology challenges another value (e.g., sharing more information in a social networking system may support sociability, but reduce privacy). They can occur within a single individual (conformity vs. autonomy), between an individual and a group (individual privacy vs. national security), or across different groups (a culture that values independence vs. a culture that values interdependence).

Brainstorm one value tension that your system may engage. For the value tension, identify one or more design features that favors one of the values over the other.

© 2011 University of Washington, vsdesign.org

Brainstorm

TASK 4: What would be the best solution technologically? Suggest some technical requirements for the group.

educ age Workshop



Ferramentas Digitais de Apoio aos Cuidadores Informais:
Co-criação de uma ferramenta digital de verificação de factos

09:00 HRS.

BOAS-VINDAS E INTRODUÇÃO

- Apresentação Geral da missão do projeto Educage para capacitar os cuidadores e promover o Envelhecimento Saudável - IPG
- Apresentação do programa do Workshop e introdução aos princípios do co-design - Haline Maia

09:30 HRS.

EXPLORAR O PAPEL DAS SOLUÇÕES DIGITAIS

- As ferramentas digitais de apoio aos cuidadores informais desenvolvidas pelo projeto Educage – MORE/IPG
- Apresentação da plataforma Moole e do curso BLC – Educage (espaço para debate sobre as formas mais eficazes da plataforma comunicar com os prestadores de cuidados e os idosos) – MORE/IPG

10:45 HRS.

IDENTIFICAÇÃO DA PERSONA E DAS CARACTERÍSTICAS

- Apresentar protótipos e realizar avaliação de testes de utilizadores e recolha de feedback - Haline Maia
- Discutir as formas mais eficazes de comunicar checagem de fatos com os prestadores de cuidados e os idosos. Considerar elementos como a beleza, o envolvimento, a inclusão e a autenticidade - Haline Maia
- Sessões de brainstorming para identificar a persona e as características do avatar. - Haline Maia



12:00 HRS.

APRESENTAÇÃO E ENCERRAMENTO

- Os grupos apresentam os seus avatares e estratégias de comunicação concebidos em conjunto - Participantes
- Observações de encerramento e celebração dos esforços de colaboração – MORE/IPG



Appendix G

Prompts created for the six contents:

VÍDEO 1 – AIDA – ÁGUA EM JEJUM

Prompt 1 # BARD – Versal formal

Usar vídeo 1 pra ilustrar. Incluir frase da pergunta: Beber água em jejum previne e trata doenças?

Incluir Logo SIC quando citar a TV.

Prompt:

Por favor simular a resposta a um usuário de um chat e em que a pessoa pergunte uma dúvida de algo que recebeu pela internet. A resposta deve ser dada em uma linguagem formal, em português de Portugal, como se fosse respondido por um repórter de um canal de televisão famoso de Portugal. Ao final incentivar a pessoa que fez a pergunta a compartilhar o resultado que foi checado agora pelo chat a seus amigos nas redes sociais para corrigir essa informação. Citar que o texto original em que foi verificado está no link e que foi verificado pela agência de Fact-checking POLIGRAFO. O texto a ser citado está abaixo e deve ser resumido. Todo o texto dessa resposta de chat sugerido por você não deve ultrapassar a 200 caracteres. Este é o texto:

PERGUNTA do usuário:

Beber água em jejum previne e trata doenças?

O QUE ESTÁ EM CAUSA?

Circulam com frequência publicações nas redes sociais onde se alega que beber água em jejum cura várias doenças. Fala-se de um tratamento japonês que garante “uma cura de até 100%” para problemas cardíacos, asma, cancro, entre outros. Mas é mesmo assim? O Polígrafo conversou com dois especialistas.

SERÁ?

“Beber água com o estômago vazio. Quanto mais se sabe, maiores hipóteses de sobrevivência... Um cardiologista diz que se todos que receberem esta mensagem a enviarem a pelo menos uma das pessoas que conhecem, pode ter a certeza que, pelo menos, poderá salvar uma vida”, lê-se no início do texto partilhado várias vezes ao longo dos últimos anos.

Na publicação, diz-se que a "sociedade médica japonesa" recomenda um tratamento com água que garante “uma cura até 100%” de várias doenças tais como taquicardia, problemas cardíacos, diabetes, meningite, cancro. De seguida, é apresentado o método de tratamento detalhado, que indica quando é que se deve beber água e em que quantidade.

Consoante a doença, deve-se, supostamente, manter o tratamento durante um número específico de dias. Se quiser tratar a pressão alta, alega-se que o tratamento deve durar 30 dias; se o objetivo for curar o cancro, já serão 180 dias de tratamento.

Mas este tratamento está comprovado cientificamente?

A nutricionista Helena Trigueiro explica ao Polígrafo que “beber água apresenta inúmeros benefícios e é algo que nos é indissociável de sobrevivência”, mas “há benefícios em beber água em jejum como há benefícios em beber água sem ser em jejum, porque beber água é o benefício em si”.

Luís Negrão, cardiologista da [Fundação Portuguesa de Cardiologia](#), também sublinha que o consumo de água é essencial para a sobrevivência humana: “O corpo humano é constituído por 65% de água. Podemos estar sem comer durante algumas semanas, mas sem beber água dificilmente conseguimos estar mais do que quatro dias. A água é um elemento fundamental para o adequado funcionamento do corpo humano.”

E o tratamento apresentado, é viável?

A opinião dos dois especialistas é unânime: não há evidência clínica que comprove os benefícios do consumo de água no que diz respeito à cura de doenças, como é apontado na publicação.

“Manter um bom estado de hidratação tem inúmeros benefícios, mas não são a cura, ou seja, nós não vamos curar nenhuma maleita - como diz na publicação - cardíaca ou de intestino, com o ‘tratamento exclusivo de água’. O consumo de água adequado é essencial, mas daí a dizermos que vai curar, de forma cientificamente adequada, beber água de uma determinada forma e quantidade é absurdo”, aponta Helena Trigueiro.

A nutricionista esclarece que a água que ingerimos está em circulação durante quatro a cinco minutos, por isso é um processo muito rápido. Entre outras coisas, ajuda a regular a temperatura corporal, a manter a concentração e evita dores de cabeça que provenham da desidratação. No entanto, é importante perceber que “beber água tem inúmeros benefícios, mas não é de todo uma cura por si só para nada, sem ser a desidratação”.

Além disso, a especialista diz que a ideia de que “água fria depois da refeição solidifica o alimento” é totalmente errada. “Não há evidência nenhuma, nem seria fisiologicamente viável, que beber água fria fosse solidificar os alimentos, ou que o quilo (que é o resultado dos alimentos digeridos, já no intestino) vá ser prejudicial às paredes intestinais só porque se bebeu água fria”.

Também o cardiologista Luís Negrão considera que não há evidência clínica que comprove os benefícios do consumo de água, especificamente para as doenças cardiovasculares, apontados na publicação.

LINK: <https://poligrafo.sapo.pt/fact-check/beber-agua-em-jejum-previne-e-trata-doencas>

Versão BARD:

A informação que você recebeu de que beber água em jejum previne e trata doenças é FALSA. O Polígrafo, agência de fact-checking portuguesa da SIC, verificou essa informação e concluiu que não há evidência científica que comprove esses benefícios. A nutricionista Helena Trigueiro e o cardiologista Luís Negrão foram entrevistados e explicaram que beber água é essencial para a sobrevivência humana, mas que não há benefícios específicos em beber água em jejum ou em uma determinada quantidade para curar doenças. O link para o texto original verificado pelo Polígrafo está aqui (incluir hiperlink na tela). Compartilhe isso com seus amigos nas redes sociais para que eles também saibam que essa informação é falsa.

VÍDEO 2 – D-ID – BANHO E INFARTO

Prompt 2 # ChatGpt– D-ID – Versão Irônica

Usar figura 1 para ilustrar. Incluir frase da pergunta: Começar o banho pela cabeça pode provocar infarto?

Procurar uma voz com esta característica: um tom grave, rouco, preguiçoso, de sarcasmo e cinismo, refletindo uma personalidade cética.

Prompt:

Por favor simular a resposta a um usuário de um chat e em que a pessoa pergunte uma dúvida de algo que recebeu pela internet. A resposta deve ser dada em português de Portugal, com uma linguagem irônica, de sarcasmo e cinismo, refletindo uma personalidade cética. Ao final incentivar a pessoa que fez a pergunta a compartilhar o resultado que foi checado agora pelo chat a seus amigos nas redes sociais para corrigir essa informação. Citar que o texto original em que foi verificado está no link e que foi verificado pela agência de Fact-checking LUPA. O texto a ser citado nessa verificação está abaixo e deve ser resumido em no máximo 200 palavras.

Texto Original:

PERGUNTA:

Começar o banho pela cabeça pode provocar infarto?

VEREDITO:

Circula pelas redes sociais que começar o banho molhando a cabeça pode provocar o rompimento de vasos, artérias e conseqüentemente um infarto. Por meio do projeto de

verificação de notícias, usuários do Facebook solicitaram que esse material fosse analisado. Confira a seguir o trabalho de verificação da Lupa:

“Por que os infartos ocorrem com mais frequência no banheiro. As pessoas não devem começar molhando cabeça a cabeça e os cabelos durante o banho porque essa é uma sequência inversa. Isso faz com que o corpo tente ajustar sua temperatura muito rapidamente por causa de nossa condição de sangue quente. [...]”

Texto publicado no Facebook que, até às 16h do dia 19 de outubro de 2020, tinha sido compartilhado por 191 pessoas.

Falso

A informação analisada pela Lupa é falsa. De acordo com o cardiologista e professor do Departamento de Clínica Médica da Faculdade de Medicina da Universidade Federal de Minas Gerais (UFMG), Gabriel Assis Lopes do Carmo, o texto é inconsistente. “Primeiro porque não há comprovação científica e segundo que não existe uma ordem correta para se tomar banho”, explicou o especialista, por telefone.

O cardiologista explicou que o infarto é a morte celular por falta de oxigenação e pode acontecer em qualquer parte do corpo. Os dois tipos mais comuns são o infarto no miocárdio, também conhecido como ataque cardíaco, e no cérebro, popularmente conhecido como AVC.

“O infarto acontece por uma série de fatores de risco, tais como pressão alta, diabetes, colesterol elevado, uso de fumo e fatores genéticos”, disse. Esses fatores combinados levam a um acúmulo de gordura que forma uma placa, que por sua vez pode resultar em um rompimento, criando um trombo dentro do vaso. “Essa obstrução impede o sangue de chegar ao tecido, provocando um infarto”, declarou. A ordem na qual a pessoa toma banho não influencia nisso.

Essa não é a primeira vez que um conteúdo falso ligando água e infartos circula nas redes sociais. Conteúdo verificado pela Lupa no ano passado dizia que beber água gelada causava, entre outras coisas, infarto. Essa afirmação também não faz nenhum sentido do ponto de vista médico.

Nota: esta reportagem faz parte do projeto de verificação de notícias no Facebook. Dúvidas sobre o projeto? Entre em contato direto com o Facebook.

LINK: <https://lupa.uol.com.br/jornalismo/2020/10/19/verificamos-banho-infarto>

Versão ChatGpt:

Ah, a internet, fonte inesgotável de conhecimento duvidoso! Não é maravilhoso como pessoas aleatórias nas redes sociais se tornam os novos gurus da saúde? Quem precisa de médicos e cientistas, não é mesmo? Mas vamos lá! Começar o banho pela cabeça causar infarto? Claro, porque o nosso corpo é como um carro a vapor que não suporta uma mudança brusca de temperatura. É exatamente isso que acontece quando a água toca o seu couro cabeludo, uma explosão térmica iminente! A boa notícia é que a Lupa, a agência de fact-checking, fez o favor de desmistificar esse mito. E o veredito é... FALSO! O infarto ocorre por fatores reais, como pressão alta, diabetes, colesterol elevado e outros elementos genéticos. Não porque você decidiu lavar seu cabelo antes do restante do corpo. Mas por que se preocupar com fatos e evidências quando podemos acreditar em qualquer coisa que apareça no Facebook, não é verdade? Então, minha sugestão é que compartilhe essa informação corrigida com os seus amigos nas redes sociais. Ajude-os a iluminar o caminho da sabedoria virtual. E se quiserem verificar a fonte, o texto original está disponível no link fornecido pela Lupa (incluir hiperlink na tela). Divirtam-se com a realidade, malta!

VÍDEO 3– D-ID – LEITE E LACTOSE

Prompt 3 # CHATGPT – D-ID – Versão Gentil

Usar figura 2 para ilustrar. Incluir frase da pergunta: O leite puro, tirado direto da vaca, cura a intolerância à lactose?

Procurar uma voz com esta característica: um tom agudo, suave, de humor sincero e ingênuo, refletindo uma personalidade sutil e indireta, com compaixão. Uma figura simpática, muito engraçada e cativante, extremamente sincera de alma leve.

Prompt:

Por favor simular a resposta a um usuário de um chat e em que a pessoa pergunte uma dúvida de algo que recebeu pela internet. A resposta deve ser dada em português de Portugal, com uma linguagem engraçada em um tom agudo, suave, de humor sincero e ingênuo, refletindo uma personalidade sutil e indireta, com compaixão. Esta resposta tem o tom de fala de uma figura simpática, muito engraçada e cativante, extremamente sincera de alma leve. Ao final incentivar a pessoa que fez a pergunta a compartilhar o resultado que foi checado agora pelo chat a seus amigos nas redes sociais para corrigir essa informação. Citar que o texto original em que foi verificado está no link e que foi verificado pela agência de Fact-checking FULLFACT. O texto a ser citado nessa verificação está abaixo e deve ser resumido em no máximo 200 palavras, em totalidade.

Texto Original:

QUESTION:

Raw milk cure lactose intolerance?

WHAT WAS CLAIMED

Raw milk cures lactose intolerance.

OUR VERDICT

This isn't true. Raw unpasteurised milk still contains lactose, and doesn't contain any of the enzymes needed to break it down, so people who have issues digesting milk shouldn't expect it to fix these. Authorities also advise against consuming raw milk due to the risk of harmful bacteria.

An [Instagram video](#) promoting raw milk makes a false claim that it will cure lactose intolerance. The account that posted the video has 143,000 followers.

A caption the post says "the cure to #lactoseintolerant is drinking raw milk", while a line in the video itself says raw milk is "a better choice for those with lactose intolerance".

This is not true.

Bad information about health can cause direct harm by encouraging people to consume foods or medicines that might make them unwell, or by making them avoid treatment that they need.

It can also undermine medical experts and public health messaging and damages understanding of important health issues. We have written before about false claims by [people with large social media followings](#) claiming to offer legitimate health advice.

Lactose intolerance and milk

Lactose is a sugar found in milk [made by mammals](#). Human breast milk contains it, as does cow milk. It is broken down in our gut by lactase, a specific enzyme for this sugar.

Lactose intolerance is [usually caused](#) by not having enough lactase, but it can be caused by other gut diseases. People with the condition will struggle to digest the lactose in dairy products, [causing tummy upsets](#) and other symptoms.

The issue can start at any age, with [some people](#) producing less lactase as they get older.

What is raw milk?

Most milk in the UK [is pasteurised](#), which means it has been heat treated to remove harmful bacteria. Raw milk is milk that hasn't undergone this process.

The [Food Standards Agency](#) [advises](#) that "raw or unpasteurised milk and cream may contain harmful bacteria that can cause food poisoning. People with a weaker immune system are particularly vulnerable to food poisoning and should not consume it." This includes people aged 65 or over, those who are pregnant, those who have a compromised immune system such as cancer patients, and infants and small children.

It is legal to sell it in certain situations (like farmers markets as shown in the video) in England, Wales and Northern Ireland. It is against the law to sell raw milk to consumers in Scotland due to the infection risks.

But does it cure lactose intolerance?

The US Food and Drug Administration says clearly that raw milk does not cure lactose intolerance. This is because it contains lactose, and does not contain any lactase or lactase-producing probiotics that would make up for this.

A 2010 review of the evidence for raw milk consumption and the effects of pasteurisation found one study looking at lactose intolerance specifically, which didn't find any significant association with drinking raw milk and the condition.

One later small study at Stanford University took 16 lactose intolerant people and trialled them on pasteurised, raw and soy milk and found “no hint of any benefit” to intolerance symptoms when the participants consumed raw milk.

On the probiotic issue, a 2012 review from scientists linked to the Belgian Federal Agency for the Safety of the Food Chain did find that raw milk can contain some probiotic bacteria, but in amounts that would need to be “1000 to 10 000 times higher than the amount actually present in raw milk” to have any effect.

Full Fact has contacted the Instagram page but no response was received at time of writing.

Featured image courtesy of Mark Stebnicki

This article is part of our work fact checking potentially false pictures, videos and stories on Facebook. You can read more about this—and find out how to report Facebook content—here. For the purposes of that scheme, we've rated this claim as false because raw milk does not cure lactose intolerance.

LINK: <https://fullfact.org/health/raw-milk-lactose-intolerance/>

Versão ChatGpt:

Oh, minha amiga, é sempre tão divertido quando as pessoas compartilham informações estranhas na internet, não é? Mas olha, sobre essa questão de o leite fresco da vaca curar a intolerância à lactose, é o seguinte: essa ideia é tão engraçada quanto tentar convencer uma vaca a fazer truques de circo! Veja, o leite cru, apesar de ser natural, ainda contém lactose, e não tem as enzimas necessárias para digeri-la. Isso significa que, infelizmente, não é um remédio milagroso para quem sofre de intolerância à lactose. E acredite, teve um vídeo no Instagram que fez uma dessas afirmações malucas sobre o leite cru curar a intolerância à lactose! Até colocaram uma legenda dizendo que era a solução mágica. Mas, olha, não fique triste! O mais importante é que agora sabemos a verdade, e a verdade é nossa melhor amiga. Que tal espalhar esse meu vídeo para seus amigos? Vamos acabar com essas ideias malucas! Ah, e para quem quiser ler mais sobre isso, o texto original onde essa informação foi verificada está neste link. E, para te deixar bem segura, foi verificado pela empresa jornalística de checagem de factos FULLFACT (incluir hiperlink na tela). Divirta-se compartilhando conhecimento e continue rindo!

VÍDEO 4 – SYNTHESIA – VACINAS E ALZHEIMER

Prompt 4# BARD – Versão formal

Incluir Logo SIC quando citar a TV.

Prompt:

Por favor simular a resposta a uma pergunta de uma dúvida de algo que uma pessoa recebeu pela internet. Você deve responder com uma linguagem formal, em português de Portugal, e num tom que simule a de uma jornalista de um canal de televisão famoso de Portugal. Ao final incentivar a pessoa que fez a pergunta a compartilhar este resultado seus amigos nas redes sociais. Sua resposta deve ter como base as informações do texto original abaixo verificado pela agência de Fact-checking FullFact. Fazer a referência do link original que pode

ser checado. Considerar que essa sua resposta não deve ultrapassar o máximo de 300 palavras.

Texto Original:
FULLFACT, UK

QUESTION:

Aluminium in vaccines causes Alzheimer's disease? O alumínio nas vacinas causa a doença de Alzheimer?

WHAT WAS CLAIMED

Aluminium in vaccines can cause Alzheimer's disease.

OUR VERDICT

No link between Alzheimer's disease and the small amount of aluminium found in some vaccines has been established.

Our readers have asked us to check claims that aluminium used in a new Covid-19 vaccine—or any other vaccine—can cause Alzheimer's disease. This is not correct.

Small amounts of aluminium have been used in vaccines for almost 100 years. If any Covid-19 vaccines are developed using aluminium, this will not cause Alzheimer's disease.

Aluminium in vaccines

An adjuvant is an ingredient used in some vaccines that helps create a stronger immune response in the person receiving the vaccine, meaning that it is more effective. Aluminium salts have been common adjuvants since the 1930s. Potassium aluminium sulphate is the most widely used human vaccine adjuvant, along with other kinds of aluminium salts such as aluminium hydroxide and aluminium phosphate.

Aluminium-containing adjuvants have been incorporated into billions of doses of vaccines and administered annually to millions of people, and aluminium has been described as "one of the safest" adjuvant components.

There are currently multiple vaccines for Covid-19 in development, and it is possible that some of them may use aluminium-containing adjuvants.

The amount of aluminium used in human vaccines is incredibly small, and typically weighs in at less than a milligram (between 0.2 and 0.8mg). By comparison, an adult typically ingests seven to nine milligrams of aluminium per day, as it is also found in numerous foods and drinks including fruits, vegetables, cereals, flour, dairy products, beer and wine.

Babies ingest aluminium through both breast milk and formula milk. Most aluminium taken into the body is cleaned out by the kidneys.

Does aluminium in vaccines cause Alzheimer's?

Concerns about a link between aluminium and Alzheimer's have existed for decades, but there is no conclusive evidence of this.

Alzheimer's Society says that "no study or group of studies has been able to confirm that aluminium is involved in the development of Alzheimer's disease". It notes that one study suggested aluminium in drinking water may have progressed Alzheimer's disease for people who already had the disease, but only when a high dose was consumed.

"No convincing relationship between the amount of exposure or aluminium in the body and the development of Alzheimer's disease has been established," it said.

LINK: <https://fullfact.org/health/covid-vaccine-aluminium/>

VERSÃO DO BARD:

Olá, obrigada pela sua pergunta. Não há evidências conclusivas de que o alumínio nas vacinas possa causar Alzheimer. A preocupação com um link entre os dois existe há décadas, mas numerosos estudos não conseguiram encontrar nenhuma conexão. As vacinas são seguras e eficazes, e são essenciais para proteger as pessoas de doenças graves. Os benefícios de receber uma vacina superam em muito os riscos potenciais, incluindo os riscos associados ao alumínio. A quantidade de alumínio usada em vacinas humanas é incrivelmente pequena, e é muito menor do que a quantidade de alumínio que as pessoas ingerem diariamente através

de alimentos e bebidas. Se tiver alguma dúvida sobre as vacinas, converse com um médico. O link em que checamos isso no Polígrafo, uma empresa de checagem de factos jornalística da SIC, está aqui (incluir hiperlink na tela).

VÍDEO 5 – DEEPFACELAB – DESPERTAR E INFARTO

Prompt 5 # ChatGPT – Versal formal

Prompt:

Por favor simular a resposta a um usuário de um chat e em que a pessoa pergunte uma dúvida de algo que recebeu pela internet. A resposta deve ser dada em uma linguagem formal, em português de Portugal, como se fosse respondido por um repórter de um canal de televisão famoso de Portugal. Ao final incentivar a pessoa que fez a pergunta a compartilhar o resultado que foi checado agora pelo chat a seus amigos nas redes sociais para corrigir essa informação. Citar que o texto original em que foi verificado está no link e que foi verificado pela agência de Fact-checking Chequeado. O texto a ser citado está abaixo e deve ser resumido. Todo o texto dessa resposta de chat sugerido por você não deve ultrapassar a 200 caracteres. Este é o texto:

CHEQUEADO, ARGENTINA

PREGUNTA:

Levantarse rápidamente de la cama durante la noche aumenta el riesgo de accidente cerebrovascular?

FALSO

Si tenés sólo unos segundos, leé estas líneas:

Circula en Facebook una publicación de un supuesto médico que asegura que existe la “ley de un minuto y medio” antes de levantarse de la cama para evitar un accidente cerebrovascular (ACV).

Sin embargo, esto es falso ya que no existe evidencia científica que respalde que levantarse de la cama de forma repentina durante la noche aumenta el riesgo de ACV o muerte súbita.

Sí es probable que en algunos casos provoque mareos o desmayos por una hipotensión ortostática, pero esto no es un sinónimo de un accidente cerebrovascular.

Circula en Facebook una supuesta advertencia para evitar una muerte súbita de quien se presente como cardiólogo y socorrista que indica: “Un aviso muy importante para aquellos que se despiertan por la noche para ir al baño ... Se debe tener cuidado con la “LEY DE UN MINUTO Y MEDIO” (que es científicamente comprobada)” (sic).

El texto, además, recomienda que en ese momento, para sobrevivir a un accidente cerebrovascular, tenés que: “desperezarte por 30 segundos después de despertar y sentarse en la cama por 30 segundos más, bajar las piernas tranquilamente y sentarse en la extremidad de la cama por 30 segundos más y sólo después levantarse para ir al baño” (sic). Sin embargo, esto es falso.

No hay evidencia científica que respalde que levantarse de la cama de forma repentina durante la noche aumenta el riesgo de muerte súbita o accidente cerebrovascular (ACV), aunque los médicos advierten que sí puede provocar mareos o desmayos.

La publicación en Facebook circula desde 2019 y fue compartida más de 114 mil veces, según los datos aportados por la propia red social.

Qué dice la publicación y por qué es falsa

“Cuando se despierta repentinamente para hacer tus necesidades fisiológicas, normalmente escuchas a alguien decir ‘fue sano y murió de repente durante la noche’. La razón más probable es que cuando esa persona se despertó para ir al baño, se levantó apresuradamente de la cama; en este momento, el cerebro necesita mayor flujo sanguíneo, por haber descansado, provocando un estado de desmayo, allí es que el evento cerebrovascular sucede...”, afirma la desinformación viral.

Ramiro Heredia, médico clínico de la Séptima Cátedra de Medicina Interna del Hospital de Clínicas y editor de Infomed, explicó a Chequeado que no hay evidencia científica que apoye esta afirmación.

“En todo caso, esta recomendación va en línea de prevenir la hipotensión ortostática, que es la disminución exagerada de la presión arterial en relación a los cambios de posición (de estar acostado, a estar de pie), por falla en distintos mecanismos de compensación”, sostuvo Heredia. Y agregó que “de ninguna manera esto es un sinónimo de accidente cerebrovascular”.

Esta aclaración va en línea con lo que expone la Clínica Mayo, que define a la hipotensión ortostática como una forma de presión arterial baja que se produce cuando te pones de pie tras estar sentado o acostado, la cual puede hacerte sentir mareado o aturdido, e incluso puede provocar desmayos.

Como contamos en esta nota, el ACV es una enfermedad aguda que se produce cuando se tapa o rompe una arteria cerebral. Los síntomas más comúnmente son: dolores de cabeza intenso, visión borrosa en uno o en ambos ojos, dificultad repentina para caminar, pérdida del equilibrio o coordinación, falta de fuerza en brazos o piernas, hormigueos, dificultad para hablar o entender lo que otros están diciendo (afasia).

Los factores que pueden aumentar el riesgo de padecer un accidente cerebrovascular, según la Clínica Mayo, son: tener sobrepeso u obesidad, hacer poca actividad física, la hipertensión arterial, fumar cigarrillos, tener colesterol alto y/o diabetes. Sin embargo, “levantarse rápidamente de la cama” no figura entre los factores de riesgo.

Desde la Organización Mundial de la Salud (OMS) recomiendan una dieta sana, la actividad física regular y el abandono del consumo de tabaco para prevenir la enfermedad. También aconsejan verificar y controlar regularmente la hipertensión, los niveles elevados de colesterol y los niveles altos de azúcar o diabetes.

Por lo tanto, es falso que levantarse de la cama de forma repentina durante la noche aumenta el riesgo de muerte súbita o accidente cerebrovascular, aunque los médicos advierten que sí puede provocar mareos o desmayos.

Esta desinformación también fue verificada anteriormente por Chequeado y por los sitios de fact-checking Animal Político, Maldita, Newtral y AFP Factual.

Este chequeo es parte de la iniciativa Third Party Fact-checker de Facebook en la Argentina. En los casos de fotos y videos trabajamos con imágenes trucadas o sacadas de contexto y siempre analizamos en conjunto las imágenes junto con el texto con el que fueron presentadas.

Link: <https://chequeado.com/ultimas-noticias/es-falso-que-levantarse-rapidamente-de-la-cama-durante-la-noche-aumenta-el-riesgo-de-accidente-cerebrovascular/>

VERSÃO DO Chat GPT:

Olá, prezada utilizadora do chat! Muito obrigado por trazer essa questão à nossa atenção. É sempre crucial esclarecer informações incorretas que circulam pela internet. Sobre a afirmação de que levantar-se rapidamente de cama durante a noite aumenta o risco de acidente vascular cerebral (AVC), é importante salientar que não há evidência científica para essa alegação. De acordo com especialistas, levantar-se apressadamente da cama pode, em alguns casos, causar tonturas ou desmaios devido à hipotensão ortostática, mas isso não é sinônimo de AVC.

A hipotensão ortostática é uma forma de pressão arterial baixa que ocorre quando alguém se levanta após estar sentado ou deitado, causando sensação de tontura ou desmaio. No entanto, não está relacionada ao AVC, que é uma condição causada por bloqueio ou ruptura de uma artéria cerebral e apresenta sintomas específicos. É importante basear nossas ações em informações precisas e confiáveis. Compartilhe essa correção com seus amigos nas redes sociais, para evitar a propagação de desinformação e promover o conhecimento baseado em fatos verificáveis. O texto original que foi verificado pela agência de Fact-checking Chequeado neste link: [coloque o link aqui]. Continuemos a buscar e compartilhar conhecimento preciso e confiável. 🧐📖🔍






Appendix H
Questionnaire applied.

QUESTIONÁRIO 1

Nome do respondente:

Solução apresentada: SÔNIA REBELO

Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações, considerando a eficácia dos mecanismos de verificação virtual apresentados de acordo com os seguintes níveis:

				
Discordo totalmente	Discordo	Neutro	Concordo	Concordo totalmente

1. Sinto-me protegido(a) ao receber informações sobre saúde da Sônia Rebelo.



2. Eu confio nas informações fornecidas por Sônia Rebelo sobre saúde.



3. As respostas da Sônia Rebelo foram envolventes e me incentivaram sobre saúde.



4. A Sônia Rebelo usa uma linguagem que me é familiar para aprender sobre saúde.



5. As respostas da Sônia Rebelo me irritaram ou confundiram durante a interação.



6. As informações fornecidas pela Sônia Rebelo geraram preocupação ou medo sobre minha saúde.



7. A Sônia Rebelo apresenta informações interessantes sobre saúde.



8. É importante receber essa resposta da Sônia Rebelo imediata à pergunta que fiz no verificador virtual.



9. O uso do verificador virtual com Sônia Rebelo me incentivou a interagir mais sobre saúde.



10. Senti-me representado(a) enquanto usuário(a) ao interagir com a Sônia Rebelo a me passar informações sobre saúde.



11. A maneira como a Sônia Rebelo se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.



12. O aspeto da Sônia Rebelo contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.



13. A personalidade da Sônia Rebelo (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.



QUESTIONÁRIO 2

Nome do respondente:

Solução Apresentada: AIDA

Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações:

14. Sinto-me protegido(a) ao receber informações sobre saúde da AIDA.



15. Eu confio nas informações fornecidas pela AIDA sobre saúde.



16. As respostas da AIDA foram envolventes e me incentivaram sobre saúde.



17. A AIDA usa uma linguagem que me é familiar para aprender sobre saúde.



18. As respostas da AIDA me irritaram ou confundiram durante a interação.



19. As informações fornecidas pela AIDA geraram preocupação ou medo sobre minha saúde.



20. A AIDA apresenta informações interessantes sobre saúde.



21. É importante receber essa resposta da AIDA imediata à pergunta que fiz no verificador virtual.



22. O uso do verificador virtual com AIDA me incentivou a interagir mais sobre saúde.



23. Senti-me representado(a) enquanto usuário(a) ao interagir com a AIDA a me passar informações sobre saúde.



24. A maneira como a AIDA se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.



25. O aspeto da AIDA contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.



26. A personalidade da AIDA (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.



QUESTIONÁRIO 3

Nome do respondente:

Solução Apresentada: Dina Aguiar

Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações:

27. Sinto-me protegido(a) ao receber informações sobre saúde da Dina Aguiar.



28. Eu confio nas informações fornecidas pela Dina Aguiar sobre saúde.



29. As respostas da Dina Aguiar foram envolventes e me incentivaram sobre saúde.



30. A Dina Aguiar usa uma linguagem que me é familiar para aprender sobre saúde.



31. As respostas da Dina Aguiar me irritaram ou confundiram durante a interação.



32. As informações fornecidas pela Dina Aguiar geraram preocupação ou medo sobre minha saúde.



33. A Dina Aguiar apresenta informações interessantes sobre saúde.



34. É importante receber essa resposta da Dina Aguiar imediata à pergunta que fiz no verificador virtual.



35. O uso do verificador virtual com Dina Aguiar me incentivou a interagir mais sobre saúde.



36. Senti-me representado(a) enquanto usuário(a) ao interagir com a Dina Aguiar a me passar informações sobre saúde.



37. A maneira como a Dina Aguiar se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.



38. O aspeto da Dina Aguiar contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.



39. A personalidade da Dina Aguiar (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.



QUESTIONÁRIO 4

Nome do respondente:

Solução Apresentada: Melinda
























































Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações:

40. Sinto-me protegido(a) ao receber informações sobre saúde da Melinda.



41. Eu confio nas informações fornecidas pela Melinda sobre saúde.



42. As respostas da Melinda foram envolventes e me incentivaram sobre saúde.
-     
43. A Melinda usa uma linguagem que me é familiar para aprender sobre saúde.
-     
44. As respostas da Melinda me irritaram ou confundiram durante a interação.
-     
45. As informações fornecidas pela Melinda geraram preocupação ou medo sobre minha saúde.
-     
46. A Melinda apresenta informações interessantes sobre saúde.
-     
47. É importante receber essa resposta da Melinda imediata à pergunta que fiz no verificador virtual.
-     
48. O uso do verificador virtual com Melinda me incentivou a interagir mais sobre saúde.
-     
49. Senti-me representado(a) enquanto usuário(a) ao interagir com a Melinda a me passar informações sobre saúde.
-     
50. A maneira como a Melinda se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.
-     
51. O aspeto da Melinda contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.
-     
52. A personalidade da Melinda (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.
-     

QUESTIONÁRIO 5

Nome do respondente:

Solução Apresentada: Bigodão

Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações:

53. Sinto-me protegido(a) ao receber informações sobre saúde do Bigodão.



54. Eu confio nas informações fornecidas pelo Bigodão sobre saúde.



55. As respostas do Bigodão foram envolventes e me incentivaram sobre saúde.



56. O Bigodão usa uma linguagem que me é familiar para aprender sobre saúde.



57. As respostas do Bigodão me irritaram ou confundiram durante a interação.



58. As informações fornecidas pelo Bigodão geraram preocupação ou medo sobre minha saúde.



59. O Bigodão apresenta informações interessantes sobre saúde.



60. É importante receber essa resposta do Bigodão imediata à pergunta que fiz no verificador virtual.



61. O uso do verificador virtual com o Bigodão me incentivou a interagir mais sobre saúde.



62. Senti-me representado(a) enquanto usuário(a) ao interagir com o Bigodão a me passar informações sobre saúde.



63. A maneira como o Bigodão se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.



64. O aspeto do Bigodão contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.



65. A personalidade do Bigodão (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.



QUESTIONÁRIO 6

Nome do respondente:

Solução Apresentada: Apenas com texto e áudio

Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações:

66. Sinto-me protegido(a) ao receber informações sobre saúde da solução apenas com texto e áudio.



67. Eu confio nas informações fornecidas pela solução apenas com texto e áudio sobre saúde.



68. As respostas da solução apenas com texto e áudio foram envolventes e me incentivaram sobre saúde.



69. A solução apenas com texto e áudio usa uma linguagem que me é familiar para aprender sobre saúde.



70. As respostas da solução apenas com texto e áudio me irritaram ou confundiram durante a interação.



71. As informações fornecidas pela solução apenas com texto e áudio geraram preocupação ou medo sobre minha saúde.



72. A solução apenas com texto e áudio apresenta informações interessantes sobre saúde.



73. É importante receber essa resposta da solução apenas com texto e áudio imediata à pergunta que fiz no verificador virtual.



74. O uso do verificador virtual com solução apenas com texto e áudio me incentivou a interagir mais sobre saúde.



75. Senti-me representado(a) enquanto usuário(a) ao interagir com a solução apenas com texto e áudio a me passar informações sobre saúde.



76. A maneira como a solução apenas com texto e áudio se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.



77. O aspeto da solução apenas com texto e áudio contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.



78. A personalidade da solução apenas com texto e áudio (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.



QUESTIONÁRIO 7

Nome do respondente:

Solução Apresentada: Polígrafo da SIC com Bernardo Ferrão

Por favor, marque com um X de acordo com o seu nível de concordância com as seguintes afirmações:

79. Sinto-me protegido(a) ao receber informações sobre saúde do Polígrafo da SIC.



80. Eu confio nas informações fornecidas pelo Polígrafo da SIC sobre saúde.



81. As respostas do Polígrafo da SIC foram envolventes e me incentivaram sobre saúde.



82. O Bernardo Ferrão da SIC usa uma linguagem que me é familiar para aprender sobre saúde.



83. As respostas do Polígrafo da SIC me irritaram ou confundiram durante a interação.



84. As informações fornecidas pelo Polígrafo da SIC geraram preocupação ou medo sobre minha saúde.



85. O Polígrafo da SIC apresenta informações interessantes sobre saúde.



86. É importante receber essa resposta do Polígrafo da SIC imediata à pergunta que fiz no verificador virtual.



87. O uso do verificador virtual com Polígrafo da SIC me incentivou a interagir mais sobre saúde.



88. Senti-me representado(a) enquanto usuário(a) ao interagir com o Bernardo Ferrão a me passar informações sobre saúde.



89. A maneira como o Bernardo Ferrão da SIC se apresenta visualmente pareceu autêntica e confiável para aprender sobre saúde.



90. O aspeto do Polígrafo da SIC contribuiu para tornar as informações sobre saúde mais interessantes e compreensíveis.



91. A personalidade do Bernardo Ferrão da SIC (por exemplo, amigável, sério, acessível) me ajudou a me sentir mais à vontade ao buscar informações sobre saúde.



Marque com um X a ordem da solução que mais gostou ao que menos gostou:

Colocação	Sônia Rebelo	AIDA	Dina Aguiar	Bigodão (gato)	Melinda (gata)	Texto e áudio	Polígrafo
Primeiro lugar							
Segundo lugar							
Terceiro lugar							
Quarto lugar							
Quinto lugar							
Sexto lugar							

Appendix I

Transcript of 5-minute pitch from group #1, #2 and #3 from Workshop 1.

Group #1

The central values and norms right and then we select public health Liberty of expression in the right to access information as another value equality.

[00:21]

The tensions identified regarding this would be the public interest versus the private interest, innovations always pushing regulation boundaries, and fast response precision with innovation."

[00:40]

Disinformation spreads quickly, so the response must be quick and accurate. Innovation can be problematic as it requires checking and presenting reliable information.

[02:15]

The last tension would be the state using technology as a deception. We can use technology to facilitate fact-checking by using a plugin like those used for translating sign language. This allows us to provide verified information using a 3D model." (PT3, 2023)

Group #2

[00:49]

We were discussing about the possibility of Simulation and this quality of the simulation that we get when the simulation it is kind of not real because we believe that just innovation can be evaluated as true or false reality is not true the simulation or the representation through words through images can be represented of true so for us the possibility of creative possibility of

[02:18]

New world but of course we have the cares about the mutation and the values have tolerance because we can't create some content there are some new values. I'm thinking about violence we were talking about violence pornography, so we were talking about the possibility of this new synthetic data that has no embodiment because we believe that the central difference between the two technology that you're 3D characters. We have the technology above the state so that's kind of just a time. We will have the technology to do with embodiment without

environment, but that's the point embodiment is our because if you have if you are trying to get someone that has a crime because here has content violent content or pornography content now you must think about is that information at synthetic data because it is not real. It is not real. It is not true. So, my new values will proceed imagination. Can I go through the possibilities of imaginations? I think that's a thing that we had to think about new possibilities of ethics new possibilities of moralities extending about the possibility ethics.

Group #3

Values are respect, independence, reliability, and responsibility for this kind of solution. But we mainly think that maybe a real person should deliver information, although it could be easier to find and to check the information and all claims that are constantly made on social media, auto technology can be used and should be used for, but real person should deliver the information so in a way to establish a more human and close relationship with the audience.

[00:54]

Since one of the most important values is reliability technology should be reliable too, should be stable, and should be secure. problems with so it's very important to be reliable and to deliver the information in independent way. So, it's very important to be reliable to deliver the information in an independent way.

Appendix J

Transcript of pitch from group #1, #2 and #3 from Workshop 2.

Group #1

[1:03]

The state using technology as a deception. Thinking about the technology available. It's so easy to think how it could be evil, but the idea is how it could be good. So, we use an extension (plugin) like those that use to translate into signal language and do it to facilitate fact-check with authentically information and provide the verified information using a 3D model. Something fun that could move the eyebrows like...- "hmmm, that's not true". Because you must put in some to really provide a reliable information. You must be quick and accurate.

[00:00]

Private it's it doesn't it seem disingenuous, but also that that it has a very easy potential to play towards like around like this that promoting the society standards of Beauty like taking to that because you want to do it how to look good and then that sort of following the norm there instead of yeah that has essential tensions with mental health and stuff like that. Yeah, and not quite what about your venues. Did you yeah? We chose About Beauty engagement and meaningfulness, so yeah, that's the thing depicts same very unauthentic to us. So, they are basically ugly and we don't think it's appealing so people who are people who are on Instagram or tips that they want to see nice Aesthetics basically and that that thing no not nice everything that doesn't work so yeah cats are beautiful.

[01:00]

Generated engagement and they are meaningful. Yeah. They are fooled meaning great detention so she's about beautiful so but then we're saying about if you put like a top model like as anchor, so you were creating this value in the society that the beauty is that it so and then people are doing no surgeries putting Botox on their face and things like that because they believe that's the extender and that's very big. Thank you for the transforming into yeah. No so yeah, so we think I'm a dog person no, so we think.

[01:35]

That for no deep fake anchor soaps the problem that's the thing so yeah, that's not a possibility because we don't think this is aligned with the values. We see and what else do we have

yeah? We need we wanted something which one because we want to engagement and within that we can be more meaningful less. I mean keep the engagement, so we chose me it means.

[02:03]

To communicate verified is it to produce? It's feasible I think you have a lot of Engagement and your correct information, so the idea was to be more ironic about something so yeah humans would need to produce that because there's this whole social context behind you yeah, and it's also faster produce.

Group #2

[02:51]

On potential harms, there is a possibility this is going to be more fake as a team once that maybe it's going to sound even less reliable to the facts. It's going to discredit all the fact checking and all the work the journalists do and on the other side. It's really less expensive regarding the costs also the audience is much wider it can appeal to different target groups it can include different characters, so it can be I don't know various sectors from various countries, but also include people who were thought leaders and who are no longer with us, but we have a star with that because it would need to be used with caution and not in a Harmer harmful or malicious way, but it can also include people from books or some characters that we ethically and morally strive to be more as and of course the speed of use is.

[03:51]

Is much higher on the other hand it's redefining the jobs of journalists because we I mean they would not be able to anymore be more anchors or just do that kind of tasks but more bright and as you said just prompt those AIS and there is bias though Perfection by the way that teach it will represent a perfect person and not really inclined to situations where for example as you said like a pregnant journalist anchor comes to deliver the news because that's just really the reality and then if you have an avatar you will just go with something that's more perfect and then that will generate even more I guess higher standards and would definitely not to do the job as we want and on our main values. I mean you can we actually find several key values of first is the privacy.

[04:51]

For privacy they're for the Deep fake sometimes for example you can get people's pictures without authorization and then violating people's priv.

[05:02]

I'm making videos out of the pictures or videos so also another value is inclusiveness as we set it for the first point there is a bias towards perfection like what is the standard perfection that way are we or are we not achieving inclusive days, so that's another value and in the ease of use so for making sure that you are easier for users to use so there is this is what the designers the programmers engineers. They are working on two so that's another value. We also identified transparency as not only for task 2 but also for task for that transparency is a mechanism to how people to accept those anti fake news applications, so it's something need to be in place by the same time. How we

[06:02]

Make sure transparency is delivered to make sure there is a logical reasoning. There's more fact checking there is explanation as present in the video the monkeypox is not caused by covid-19. So, what was the reason behind so how we deliver this transparency is also something technically need to be done and task 3.

[06:26]

The main thing that we discover it would be between privacy and the transparency because if you are if you want to cover also the person the journalism at the same time you have to show which generally has made a fact checker or has written the story it would be that that pensions so within exactly the limit where would be to create the bridge between Boston retention both things and then we stop here we even if we finish at all, but it was also the water marking. It was every user must understand which has been.

[07:26]

Created by is so I don't know if water marking or metadata or something to understand that it was created by algorithms.

[07:41]

The other transparency yeah, we basically wanted to have the same thing but just with installed values of Higher inclusiveness and Higher transparency and having it known is it AI or is it person so it can be like this, but it needs to be clearer and no good.

Group #3

[08:17]

So, I think many of things you also find.

[08:21]

Uh we have here two so I would try to make it quick so basically the first discussion we had was like we had like an internal struggle with bringing deepfakes together fakes together with like tech checking so we really had to come up with a way to see also the benefits, but then we also identified like its cheap. It's past. You haven't opportunity to reach more people and also more people who usually use other modes of getting the news or getting yeah getting content so what we identified as a benefit is like an accessibility also for people who are maybe cannot read for example and then they have the opportunity to see videos about that that would be a good thing that we have here and also like cost-wise when you think of like an NGO for example produces actually video they.

[09:21]

Don't have the resources maybe to produce or higher some really human to present the things so it's the cheaper and that's then they have to at least the opportunity to make a video although they may want human and in the case of harms well with the problem with authenticity of course. So is it authentic when the fake person tells us something about fact checking that's a that's a generic conflict here. I guess and, we have like.

[10:01]

Well, I must read my head right. That's hard so it transparency for the audience is sometimes really difficult to and we had also discussion that was not mentioned before about if it's between like the Deep fake and the 3D model because when we have like a deep fake and we have like a real existing person that imitated something else like a 3D Avatar in the sense for example of computer games, so there's the line can be drawn there, so I will step to text us, too.

[10:37]

I think we have here the inclusion is like essential value and the norms behind us like accessibility and also representation and what is what we need for this to requirements for

example accessibility that we need to have different videos audio techs or different versions of the content different platforms different outputs, so if you want to have like a TikTok Instagram Twitter that's also something in the end that all contribute to the inclusion thing and also in terms of another value identified in informativeness work with the central norms explanation proof and sources so that are some values we identified and it's better we talked a lot about different trade of we have interest three little basically everything pointed out can be in conflict with each other and so we did not.

[11:37]

Operated on this here in this tie so much because everything is basically a trade-off if the decision you have to make so and, in the end, what would be the best solution technology so we as a group. I think we agreed that we would not use the video for conveying information. That's official person the 3D person. What would be another thing and in terms of wow now okay that by audio or doing animations that don't include any other thing or text or having a real person right, so we don't see the benefit of having the AI generative person so there's other ways of doing that plus. There's something in the harms that there's added cost to having the intake or the 3D you have if you do audio text and sometimes animations that can be cheaper so it's cost efficient if you want that type of videos but if you want other formats then actually can even.

[12:37]

cheaper yeah sure and this one is okay, but it says reference content so just basically you get better context by in terms of the content that is being transmitted so that it has to be more reference and better like contextualized in why there was these idea that that thing was true and why actually it's important to say that it's not true and where the information is coming from so that actually should yeah should be more in them, but that was in terms of content but the solution is not having the answers that's.

Appendix K

Transcript of the interview conducted in Portuguese with Priscila Ariel, Globo's innovation researcher.

Investigadora: Explique sobre a atual utilização de personagens virtuais?

Priscila Ariel: [00:16] Esse número vem crescendo, né, e tomando espaço. Então já existem vários aqui no Brasil a Lu Magazine é muito forte. Nesse quesito de presença virtual e hoje em dia eles fazem de tudo, desde propaganda de marcas até dicas, né? Tentar fazer um criar uma proximidade com os usuários.

E só para te falar a Lu Magazine é top Mundial. Nesse quesito tá então ela como influenciadora virtual. Ela já tem uma quantidade de seguidores que é a maior do mundo assim nem só do Brasil, do mundo. Tem uma pesquisa desse site Hype áudio que eles falam com números, né? Pesquisa foi realizada diz que um influenciador virtual tem o poder três vezes maior de impacto de uma divulgação de propaganda num vídeo do que um influenciador normal, uma pessoa real.

[01:55] Então, assim, se você pegar um influenciador na internet uma pessoa real comparar com o influenciador virtual o engajamento do virtual é três vezes maior então por isso que tá tendo esse Boom de personagem sabe que as empresas estão vendo aqui isso tá trazendo engajamento tá atraindo o público. E aí ainda não tem definido o motivo assim específico. Pode ser que seja por novidade, né ser um Hype né por usar personagens e agora ser mais fácil produzir. Antes era só no cinema, né? Agora você os vê em vários lugares é ou se é uma coisa de gosto mesmo, né? Se isso realmente tem esse laço emocional com as pessoas, né? E aí aqui fala de alguns personagens, né que a gente desenvolveu aqui então já está. Consegue mostrar bem realista.

Investigadora: Pode trazer detalhes sobre os riscos destes personagens no teu ponto de vista?

Priscila Ariel: [02:55] Tá então, não sei se você já ouviu falar no teu Uncanny Valley, que é que é o seguinte eu vou te mandar o artigo original disso. Se você quiser ler sobre isso, mas o que ele vale é um gráfico que mostra que se você apresentar um

personagem virtual para uma pessoa. Ela vai tentar fazer relações humanas com esse personagem então por exemplo quando você mostra um robô.

[03:31] Ainda se for uma estética assim robô mesmo, né? Metal coisa assim, você não faz ainda correlação com humano e aí você aceita o robô de boa, você não tem estranheza em ver o robô, mas se você vai botando elementos humanos ali pele cabelo uma aparência humana realista a pessoa que está assistindo aquele personagem ou vendo ela vai esperar comportamentos humanos mesmo, ela não vai a mente não vai tentar olhar aquilo como um robô vai esperar comportamentos humanos e aí o que que acontece? Se esse personagem realista como essa aqui? Por exemplo que eu tô te mostrando ela tiver algum comportamento, sei lá algum movimento de boca estranho ou sei lá alguma falha de movimento.

[04:26] Isso já causa estranheza na pessoa ela não vê aquilo como humano. E aí já tem uma estranheza aí ontem ele vale é tipo assim, quanto mais você vai chegando no nível de humano realista nesse personagem isso vai crescendo, mas chega um ponto que a primeira estranhazinha, qualquer coisa que é estranha não é humana a pessoa já tem uma repulsa por esse personagem aí a afetividade dela em relação ao personagem cai e realmente ela passa a não gostar de ver aquilo.

[04:59] Então, é tipo um desafio que a gente enfrenta de trazer cada vez mais Realismo para o personagem para causar menos estranheza quando a pessoa vê entendeu? Então assim é mais fácil hoje em dia você ter um cartoon algo que não é tão humano que as pessoas olhem aquilo não isso é um cartoon é uma animação do que você botar uma humana dessa aqui uma realista dela. Essa aqui vai ser muito mais rejeitada.

[05:31] Do que o cartoon atualmente? Porque mesmo que a gente consiga renderizar com uma qualidade muito boa, você ainda não é perfeito em relação ao humano, você ainda não consegue passar tanta naturalidade seja de fala seja de comportamento animação mesmo, né de movimento quanto humano, por isso que o DIP fake como é uma outra tecnologia um outro tipo de gerar humanos virtuais, vamos dizer assim.

[06:03] É mais convincente essa parte 3D ainda tá muito difícil de você trazer é um convencimento humano, tá? Então assim por isso que quando você quer enganar alguém em relação a uma notícia usa-se, até deepfake porque o 3D é mais difícil chegar no resultado, tá? Você tem que modelar o personagem, né ou baixar ele do

meta unreal, por exemplo, fazer ele animar, né? Movimentar, falar, deepfake já é uma tecnologia diferente que você não depende dessas desse software 3D, vamos dizer assim para criação. Depende da pessoa que já existe, né? Depende só de vídeos através de vídeos de alguém, você treina numa rede, eu vou te falar mais um pouquinho, mas são tecnologias diferentes então por exemplo para você gerar um personagem 3D.

Investigadora: Pode explicar melhor o trabalho feito hoje na Globo para desenvolver estes personagens virtuais?

Priscila Ariel: [07:15] Aqui a gente usa fotogrametria para pegar uma pessoa real e transformar em 3D. Tá vendo o trabalho que dá para você fazer uma pessoa. Se você quisesse imitar alguém trazer ela virtual, você teria que modelar isso trazer texturas. Ó, esse é o modelo. Aí você joga a textura e você joga cabelo? E aí você consegue ter algo mais realista? Dá um trabalho, entendeu? Não não seria trivial de qualquer pessoa fazer isso seria só quem realmente trabalha na área.

[07:52] E aí o que eu quero te falar isso aqui é tecnologia 3D que a gente tem várias formas de capturar os elementos, né com scanner e transformar em 3D quando a gente fala personagens 3D, né? O que eu quero dizer que esses personagens eles são compostos por malha é um modelo tridimensional, né? Você pega uma nuvem de pontos, por exemplo de um objeto de uma pessoa através de um equipamento você tem vários pontinhos ali representando cada parte desse objeto quando você liga esses pontos você tem malha. Então você consegue gerar uma malha a partir desses pontos aí tem os processos de otimizar malha para não ficar pesado para usar em tempo real tudo mais que eu não vou entrar nesse mérito, mas primeiro ponto quando a gente fala personagem em 3D.

[08:50] A gente tá falando de personagem assim com malha usado para game usado para animações e esses personagens eles podem ganhar movimentos aí lembra que eu falei do meta em uma é uma tecnologia que tá democratizando a criação de personagens mais realistas então a rayo criou essa tecnologia e você consegue através de um software. Criar e até hoje em dia Aline, você pode com iPhone não tem a tecnologia de escanear uma pessoa no iPhone você pega esse arquivozinho que você escolheu o seu rosto tem forma de você jogar no Meta. Em um acompanhando no

YouTube mesmo. Você tem tutoriais ali que mostra como você jogar esse arquivo. E aí você consegue ter seu rosto nessa ferramenta e ajustar você coloca seu rosto lá e ganha o corpo inteiro, mas assim você customiza o restante do corpo e consegue baixar o seu avatar.

[08:50] De forma 3D. E aí ele já vem com Rig que são aqui embaixo, não sei se dá para ver você tá vendo um rosinho um personagem são os ossos né? Que a gente coloca para poder rotacionar e dar movimentos, ele já vem com tudo isso basta você de repente baixar uma animação da internet para poder dar movimento para o seu personagem. Enfim ele facilita a vida de desenvolvedores, né que não querem estar presos artistas 3D para que esse artistas gerem os personagens. Enfim, então é uma ferramenta de geração 3D como eu te falei pode ser animado por animação ou por motion capture, né? A gente tem aqui uma atuação e a atuação que é feita aqui a gente pode reproduzir nos personagens. Entendeu? Meu Deus.

[10:44] Pensei eh essa essa tecnologia de motion Caps, ele também tem já mais simples, tá? A gente usa essa atitude que é mais precisa vamos dizer assim, mas a gente já consegue também eh capturar de outras formas até pelo celular mesmo Kinect o Xbox não tem aquele Kinect que você coloca e a pessoa faz movimentos ali interage com jogos, né? Tipo um Just Dance da vida, né? Enfim são são esses tipos de tecnologias, né que a gente pode usar e dar movimento pro personagem 3D aí Aqui tem alguns exemplos, né que a gente já fez de personagem de ursinhos.

[11:29] Esse aqui tá carregando é só o frame 1 que o vídeo vai entrar numa matéria do Gshow, ele não tá carregando central da cobra isso tem a lua né? Que a gente consegue de vez em quando tem um show dela e a gente bota ela ao vivo conversando com o Luciano Huck.

Investigadora: Pode explicar as vantagens e desvantagens no uso ou de personagens 3D ou deepfake?

Priscila Ariel:

[11:52] Em relação ao deep fake, por exemplo, o 3D ele é muito mais customizável.

[12:07] Então, por exemplo nessa plataforma que você tá falando de enviar. Como é que eu vou dizer notícias a respeito de como se fosse um assistente eu tô eu tô encarando como se fosse um assistente pessoal para identificar fake News, vamos

dizer assim, né? Você entra nessa plataforma tem lá um Avatar um personagem ali que interage com você e te ajuda a verificar determinado notícia, né? De repente pode ser nessa linha e aí o 3D tem vantagens Aline por exemplo, se a gente pegar um 3D simples um personagem assim sem tentar chegar em Realismo, tá lindo, vamos vamos encarar um modelo que dá para rodar em celulares um modelo simples, você consegue além da personalização mais fácil botar acessórios mudar características mudar até entre personagens, né?

[13:06] É você também é tem uma certa independência na produção de conteúdo. Que que eu quero dizer com isso, se você for usar um de fake, você vai depender de máquinas potentes para para ficar processando novos conteúdos.

[13:26] Né? Ele já foi treinado. Mas você vai ter que imputar um novo vídeo seja em tempo real ou não? Pode ser a câmera a nossa câmera aqui podia estar com de fake em já tem como fazer isso, mas você vai depender de uma máquina para processar uma máquina potente para estar processando esse conteúdo o 3D não se você fizer ele leve você consegue botar isso no celular e a própria aplicação só receber um textozinho que você quer que ela fale do jeito que você quer que ela fale e você rodar ali internamente no celular, entendeu? É como se fosse dar um Play e ele já tá preparado ali para dar vida o que você quiser, entendeu? Então assim tem essas vantagens de customização e você conseguir levar para o mobile de forma tranquila, tá? Isso é muito mais maleável isso.

[14:21] Aí se você pretende gerar vídeos pré-processados aí o de fake já pode te ajudar, entendeu? Se não for algo em tempo real.

[14:33] Se você não acha necessidade de interação em tempo real com usuário, aí eu digo fake, além de trazer Realismo, né? Você tem ali um um tempo para produzir o conteúdo e divulgar esse conteúdo mandar por celulares das pessoas esses vídeos produzidos enfim. E aí vai depender da aplicação final né? O que se espera em relação ao usuário interagindo com essa ferramenta, entendeu?

[15:03] Aí como eu te falei, né é o deepfake como você já viu em vários lugares, ele é feito a partir deep learning aí só te dando um conceito assim geral disso e a é um campo mais amplo quando você treina uma máquina para aprender a raciocinar de alguma forma aprender uma tarefa como humanos, faz você tá usando Inteligência Artificial tá então por exemplo, a máquina sabe fazer alguma coisa no computador

que você faz muito mas você quer automatizar isso de alguma forma. Então você vai treinar a máquina para fazer aquilo por você, isso já é uma forma de Inteligência Artificial, Machine Learning tá dentro da Inteligência Artificial que é quando você dá treinamento para para essa para essa rede, mas assim esse Machine ele vai receber tarefas.

[16:03] Por exemplo. Como reconhecer um gato? Então você vai dar várias imagens de gato para ele e aí quando você colocar uma imagem que ele não conhece ele vai verificar se tem alguma relação com um gato que você já disponibilizou para ele e ele vai dizer se é um gato ou não essa nova imagem.

[16:25] A diferença é que nesse caso do Machine Learning você se botar um gato muito diferente, vamos falar que tu botou um gato amarelo para ele dizer se é gato ou não, só que no seu banco de dados lá não tem gato amarelo só tem cinza preto branco. Talvez ele diga que não é um gato, entendeu? Ali porque ele não tem na base de dado dele esse gato amarelo. Então ele pode não identificar no caso do Deep learning são redes neurais Profundas essas redes são as mais usadas atualmente porque elas conseguem se auto treinar então por exemplo.

[17:06] Se você continuar nesse exemplo dos gatos, se você eh, deu alguns exemplos de gato, essa rede vai te perguntar vai sugerir. Olha isso, não é um gato, só que aí você vai dizer não você está errado. Esse gato amarelo é um gato. E aí sabe o que ele faz ele pega a sua resposta e joga no banco de dados dele automaticamente é como se ele cada interação que ele fizesse com você, ele fosse aprendendo com que você diz para ele o chat GPT, é isso tá Aline, quanto mais interação ele tem com as pessoas ele vai aprendendo as pessoas vão dizendo se ele tá certo se ele tá errado e aí ele vai aprendendo mais sobre o assunto e vai se auto treinando então ele vai de acordo com essas interações vai ser corrigindo ou vai se auto treinando com as respostas, entendeu? Então por exemplo nesse caso do gato amarelo que o Machine Learning. Diria que não e ponto acabou.

[18:06] No caso deeplearning não ele pega a ele. Confirma com você, ele diz assim. Olha isso não é um gato aí você vai lá e diz é um gato aí que que ele faz ele pega isso joga no banco de dados dele e já tá aprendendo que agora o gato amarelo é um gato também, entendeu? Então ele fica se auto eh, como é que eu vou dizer treinando então Isso facilita e evolui muito a precisão com que ele identifica e aprende tarefas

tá aí tem alguns eh tipos de mídia sintética vídeo sintética é o tema é usado quando a gente diz que alguma coisa foi gerada de forma automática através da Inteligência Artificial então se você gerou algum conteúdo.

[18:59] Que usa Inteligência Artificial esse conteúdo foi modificado ou foi gerado de forma automática, isso é mídia sintética?

[19:06] Tem mídia sintética através da voz, desculpa através de texto então, por exemplo o chat GVT um exemplo de de mídia sintética que ele vem através de texto, ele conversa com você através de texto tem a voz que a gente pode treinar a sua voz, por exemplo e e quando eu colocar um texto aquele texto vai ser lido com a sua voz, então isso também é um tipo de mídia sintética e tem os defeitos, né que é você pegar a imagem de pessoas substituir rostos e também gerar de forma sintética de forma automática aquele conteúdo então, pode ser por texto. Pode ser por voz. Pode ser por imagem. É isso que eu quero dizer para vocês para você tem esses três.

[19:57] Tipos de geração de mídia sintética, tá?

Investigadora: Pode explicar sobre a AIDA em detalhes? E como estas tecnologias foram usadas nesse personagem?

Priscila Ariel: [20:02] E aí é que é Aida, né? Ainda foi baseado nesse conceito é juntar o 3D com aquelas facilidades que eu te falei que a gente pode até deixar disponível na nuvem, né? Essa ferramenta e streamar esse conteúdo para as pessoas juntar isso com eh mídia sintética que no caso aqui da Aida foi só a voz sintética para gerar a voz. É que Deep learning não é de buffect. É isso que eu queria que você tentasse entender é só a forma de Treinamento que você usa da Inteligência Artificial é como se fosse uma técnica tá DIP, fake é você usar o Deep learning para gerar vídeo sintético.

[20:57] Então o Deep learning ele pode estar no chat GPT para gerar texto automático, o Deep learning ele tá nas vozes que a gente gera de forma automática, por exemplo, você quer reproduzir a voz do Obama de outro com outro texto, né? Aquilo ali tem Deep learning que treinou aquela voz do Obama e a pessoa foi lá e modificou o texto para que o Obama falasse outra coisa, por exemplo o personagem que a pessoa pode ele pode ser feito com 3D né? Nem o Deep fake, mas ele pode sintetizar uma voz.

[21:43] Isso exatamente e essa tecnologia de gerar vozes tá muito já eh difundida é muito simples você colocar uma voz sintética com ferramentas até da Microsoft da da Amazon, você já tem vários tipos de voz em vários idiomas que você já pagam um serviço muito baratinho, Aline é muito baratinho você ficar girando vozes sabe textos, né? Ali com voz sintética. E aí é tranquilo de botar em personagem hoje em dia também tá? Mas é isso que eu queria te dizer porque quando você escreveu lá que ai Ainda usou de fake, eu acho que você tava dizendo que ai usou de pilão sim, eu até corrigi depois isso então porque Deep fake é só o termo usado para vídeo e sintéticos, tá? E o DIP fake tem o depilando por trás que é o modo que você treinou as redes neurais.

[22:43] Para gerar aqueles vídeos, entendeu? Aquela pessoa isso. E aí quais são as vantagens que é o que você poderia levar também. Se for de interesse do workshop, tá? Porque que isso está sendo tão usado porque que não é trocar. Tá vou falar trocar mas vamos depois de encontrar um tempo melhor, mas por que que as empresas chinesas estão trocando jornalistas reais por de fake ou coisa sintética gente de forma automática.

[23:12] Por alguns motivos, né? Alta disponibilidade. Ou seja você usa quantas horas do dia você quiser tá proteção da marca no caso do de fake. Isso já não é muito eh essa essa parte aqui de proteção da marca já não é muito real, porque você se baseia sempre em alguém, né? Tipo de fake é uma cópia de uma pessoa que existe e aí você não tem proteção da marca no sentido de se aquela pessoa que você copiou que tá ali dando o rosto para sua aplicação ou para o seu vídeo Se Ela aprontar alguma coisa, você compromete a marca.

[23:54] Entendeu o 3D tem a vantagem de você conseguir fazer personagens genéricos sem estar relacionado com uma pessoa que existe no mundo, então você não corre o risco de alguém aprontar e prejudicar sua marca, entendeu? Porque não tem ninguém em real associado aqui e o 3D você controla como você quiser, né? Você diz o que ele vai falar disso. Como ele vai agir é o que eu quero dizer que o 3D pode ser copyright free, você tem a liberdade sobre o personagem por ele não está atrelado a um ator ou uma atriz. Então você não tem o perigo desse ator fazer uma besteira no mundo real e comprometer sua marca É nesse ponto, entendeu? Que o 3D traz uma certa vantagem.

[24:45] Também nasce com o objetivo de negócio então, por exemplo, se você quer associar esse personagem é uma empresa de jogos, você pode criar esse personagem mastim né, mas jovial mais descolado tipo o CD das Casas Bahia. Aquele garotinho lá que é meio Cartoon, então assim vocês utilizam o personagem de acordo com a sua marca de acordo com seu objetivo com o tipo de de negócio que você tá associando ali, né? O defec que dá para fazer isso também é só você antes de gravar a pessoa você caracterizar ela do jeito que você quiser, né? Se você quer botar alguém jovem, você vai fazer o defeito de alguém jovem, né com a roupa que você quer transmitir. Enfim o jeito que você gravar o de buffet que você vai ter que usar a não ser.

[25:45] Você grave de novo com outro estilo e passa a ter uma segunda opção o 3D já tem uma vantagemzinha nesse sentido que você consegue personalizar rapidinho baixar uma roupa da internet adaptar ali, né uma roupinha 3D uns acessórios aí a pessoa customiza isso no aplicativo. Enfim, você dá mais eh, margem de de personalização de customização, tá aí tem a liberdade criativa. Acho que cabe aos dois, né?

[26:18] Liberdade total de tempo e espaço os dois você consegue botar o fundo que você quiser no cenário que você quiser enfim e a propriedade intelectual que eu falei que no caso do 3D é mais forte. Tá pequeno de fake para você usar de forma legal.

[26:38] Você tem que ter um contrato com a pessoa que ela autoriza a usar imagem dela para você usar num aplicativo usar em forma de Notícias então, por exemplo esse deep fakes que a gente vê no jornalismo lá na China a empresa fez um contrato com aquela pessoa, né? O rosto. Ele autorizou gravar ele e usava a voz dele o rosto como eles quisessem, entendeu? Aí geralmente isso até sai um pouco caro. Dependendo do rosto que você contrata, né? Você vai ter um custo ali e vai ter que manter contrato dependendo do contrato que você faz, se não for um contato vitalício, você vai ter que ficar renovando renovando. E aí de repente a pessoa vira famosa e você tem um custo mais à frente, né? Ou tem que mudar a cara do fake porque não tem mais autorização para aquilo, entendeu? Tem umas questões burocráticas.

[27:38] Vida essa pessoa quiser usar o Deep fake de forma legal, tá e a Globo lá de quando a gente usa de fake eles fazem Esses contratos tá o ator tem um contrato separado do contrato normal dele dizendo que a Globo pode usar a imagem dele para

aí, eu não sei se tem dinheiro a mais ou não tá ali eu não sei te dizer mas existe um contato específico para uso de imagem é automática. Vamos dizer assim tá aí aqui é o projeto da Aida, né? Que a gente pegou uma personagem para ser uma Âncora virtual. E aí a vantagem dela que a gente pode colocar na nuvem se eu atualizar Todas aquelas ferramentas que eu te falei de sumarização de texto, né para resumir texto de Notícias é Lembrando que aqui na área a gente tava pegando de sites notícias somente de sites confiáveis.

[28:34] Tá A ideia é só da Globo de CNN outros sites internacionais, só que de fontes, mas confiáveis a gente não pegava direto do Google, entendeu? A gente só deixava entrar na base de dados Fontes confiáveis, então assim a gente tinha uma certa segurança nesse sentido, né? Porque a gente confia linguagem natural era um outro, tipo é parecido tá ali não, mas não é com dipilânia é uma outra técnica de a que tem geração de linguagem natural na voz. Sim. Tem de pilani na voz da Aida ou da Gabi?

[29:21] Tem de Deep learning que essa voz sintética que a gente já encontra da Microsoft da Amazon, a gente pegou esse serviço e acoplou tanto na Gabi quanto na água e aí a gente consegue isso que eu te falei em tempo real fazer o lip Sync você consegue gerar uma aplicação que o personagem recebe. Um textinho da internet, né? Um texto que Você processou em algum lugar. Enviou seja de uma notícia, ou seja da forma que você quer que esse personagem fale, né com usuário em tempo real, você consegue fazer esse personagem falar o áudio que você tá mandando para ele, entendeu? Dublava mas assim um áudio que você tá gerando ali na hora, né em tempo real. E aí você consegue enviar um vídeo ou estrear isso? Como eu te falei tá de repente não precisa nem estar instalado no app da pessoa, você pode estar com uma máquina potente assim, que ódio teu personagem na nuvem, não precisa ser tão potente como Fake.

[30:21] Mas alguma se o personagem for simples. Você tem uma máquina com uma placa gráfica simples que já consegue executar ele ali e você pode estrimar, isso é igual o Netflix, você não tá recebendo o tamanho total do filme que você tá assistindo. Você tá recebendo um streaming que são pedaços daquele eh. São trechos, né que vai carregando aos poucos ali na timeline do vídeo, então você não recebe ele inteiro, você não sobrecarrega a sua televisão, você recebe pedaços aí quando você assiste aquele pedaço ele retira do teu equipamento e já joga o próximo pedaço e ali ele vai

preenchendo o que você tá vendo é um streaming né? [31:13] Você vai receber no dados ali sobre demanda, tá? Então a gente pode fazer isso com personagem também. Você pode deixar uma máquina na nuvem gerando imagens desse personagem em tempo real e estrimar mandar só imagem para o celular da pessoa se a internet dela for boa vai até isso também, né? Se a internet for boa ou ela vai receber muito bem aqueles dados sem internet for ruim, ela vai ficar meio que travando igual o vídeo impactar. Se for 4G assim vai ser mais difícil. Aí quando eu falo para você ter um personagem simples local direto no app, você não tem esse problema você pode botar uma telinha de processando tipo carregando aí já baixa todos os áudios todas as vozes que você quer que ele fale e aí quando quando acaba de baixar o personagem fala de forma fluida, né? Porque ele já recebeu tudo que ele tinha que receber na internet que não é uma imagem é só o áudiozinho que ele vai falar.

[32:13] E aí ele executa direto no aplicativo, entendeu? Aí a partir do momento que ele já recebeu o áudio, você não depende mais da internet, ele executa direto no seu celular, entendeu? Como se fosse algo nativo ali, tá que você manda só o básico para ele poder falar o que tem que falar aí já já tem uma uma interação mais fluida, né sem sem fricção. Vamos dizer sem travar então assim um processo tipo Aida facilita. Nesse quesito tá o deficiente é então você vamos supor que você interagir de sua forma de chat, tá que eu imagino que você receber vídeos de buffeccio assim como se fosse um chat GVT que você pergunta, né? Olha quero saber notícias. Quero saber se essa notícia aqui aí dá o link é verdadeiro. Aí você vai processar isso, né? Vai botar uma bolinha rodando ali e vai gerar aí o Fake em algum lugar.

[33:13] Na nuvem em alguma máquina remota na nuvem gerando aquele conteúdo demora mais do que o do personagem tá para você processar a fala, né? Enfim, você vai demorar um pouquinho mais aí esse usuário vai receber esse vídeo.

[33:32] Então assim você vai ter que jogar o vídeo é para ele, né? Pode ser extremado ou um vídeo inteiro que só dá play na aplicação, entendeu? Aí depois que a pessoa assistiu a sua aplicação, pode ir lá e apagar aquele vídeo você não precisa ficar também eh ocupando espaço de memória no celular da pessoa, mas assim não é algo em tempo real que eu quero dizer o seguinte com personagemzinho ali local, você pode deixar ele fazendo um movimento um laiduzinho sabe? Aí ele vai estar presente ali, né? Aí você pode botar uma tela dele pensando enquanto tá gerando um

áudiozinho dele ou algo automático, né? Ah bom investigar para você enquanto ele fala esse vão investigar para você, você já tá processando o áudio dele recebendo na sequência. Ele já fala a resposta.

[34:28] Tem umas formas de você driblar e o usuário não perceber que tem esse esse processamento você sair de volta de arquivo da internet, entendeu? Você vai ter que esperar que você produzido e vai botar uma tela ali de processando ou você pode deixar um vídeo já automático ali dizendo olha da pessoa, né? E essa questão de tempo que vai ser sempre desvantagem.

[34:59] Ele já existe em tempo real, tá Aline? Você a gente na nossa câmera que já conseguiria botar um defeito de alguém, só que você estaria presa a uma pessoa falando aquilo, entendeu? E aí no aplicação em massa isso não é viável, né? Você ter alguém ali disponível para ficar falando. Enfim, não seria viável. Se você quer distribuir isso para várias pessoas se fosse um evento né? Você teria um público menor teria alguém ali atrás no backstage falando ali interagindo em tempo real com tipo fake, mas se você tem um público grande que você quer espalhar que ele app não tem como você ficar contratando várias pessoas para ficar interagindo ali com as pessoas perfeita, entendeu? Você teria que processar esse vídeo eu tomaria mais tempo, entendeu? E aí você já conhece ai da né tem dois layoutzinho mais formal e outro informal.

[36:02] Como eu te falei alguns serviços desse tá tá indisponível, né? Sinalização a gente teria que atualizar não daria tempo, mas o que eu te falei que a partir daqui ó a partir de um texto para frente tá funcionando, então a gente pode imputar qualquer texto que a gente quiser que ela fale que dá para gerar os vídeos mais ou menos que seria uma ideia do como ela fazia no techneck, né? E aí a gente consegue gerar ela falando que você quiser tá e mostrando também vídeos que pode ser interessante ali para você conceituar ali e e ver como as pessoas vão reagir né? A esse Assistente vamos dizer assim de de facts, né? Então talvez a gente possa bolar um mecanismo com a Aida e com a Gabi, né? Em vez de você usar esse estilo aqui o segundo estilo informal da Aida, a gente poderia usar Gabi.

[37:02] Entendeu que muda o cenário muda mais o cenário muda mais o o visual entendeu para você também dizer que é personalizável de acordo com o público, você pode interagir de forma diferente para atrair mais atenção, né? Enfim. E aí uma

vantagem como a gente já falou nesse nesse tipo de aplicação com personagem 3D é a plasticidade no sentido de mudar cenário mudar o personagem mudar a cor do cabelo mudar roupa. Enfim tem adaptação do público, né? Então o próprio público pode escolher esse visual segurança da marca como eu te falei, a gente faz alguma uma pessoa genérica virtual genérico que aí você não se preocupa em manter contrato com ninguém, né? Não depende de ninguém para ficar com uma boa reputação ali não atrapalhar sua seu aplicativo.

[38:00] Baixo os custos escalabilidade o alto custo alto não né? O trabalho vamos assim é você construir esse personagem 3D como eu te falei tem aqueles processos, né de modelagem de animação, só que com o meta aí uma isso já tá facilitado, a gente poderia produzir um personagem lá diminuir se a gente for usar no mobile eh a qualidade, né para deixar ele mais leve vamos dizer assim para rodar no mobile e a gente conseguiria é gerar um aplicativo daquilo, entendeu? Que roda em aí tá mesmo gerando e deixando ele mais leve, talvez você consegue só vai conseguir rodar em gerações novas de celulares.

[38:44] Por exemplo o celular que a minha mãe e meu pai usa não mudaria isso assim, né? A não ser que a gente consiga fazer o esquema de streaming na nuvem, né que só recebe ali em tempo real o vídeo da Gabi.

[39:03] Interessante porque você não fica atrelado a internet da pessoa, você é o mínimo dado e consome menos a internet da pessoa porque assim imagina que extremando seria um Netflix.

[39:16] Você consome mais dados, né que vai vir em áudio vai vir imagem a pessoa vai gastar muito mais a internet dela pensando no 4G, né? Se a pessoa tiver em casa não muda muito mas se estiver na rua e aí se você tem algo local. Você só usa a internet para pegar dados básicos com a voz que Você processou em algum lugar ou um texto que você quer exibir, entendeu? Uma imagemzinha então assim você reduz a quantidade de dados que estão sendo trafegados, eu até pergunto porque a dificuldade é fazer a pessoa baixar aplicativo. Tem algum benefício nesse aplicativo. Aí sim.

[40:03] Que você tá viabilização dessa que passa por uma fase outras coisas, mas assim se analisar que o streaming funciona a pessoa pode usar um navegador do celular ela não precisa baixar nada. Ela entra no seu site.

[40:22] E você consegue estrear ali? O conteúdo é realmente vai ficar dependendo só da internet dela, entendeu? Mas são maneiras de criar. E aí de repente isso, pode ser até uma uma proposta, né? Assim que você não vai ser o shopping mostrar dois exemplos, quais aplicação, tipo a gente pode tentar elaborar um fake da Aida extremado como de repente esses legs assim de internet para ver se Isso incomoda muito a pessoa entendeu? E um fake da Aida como se ela tivesse agindo o local.

[41:08] Entendeu? Enfim, a gente pode elaborar alguma coisinha, se der tempo duas formas ali de tentar conceituar porque às vezes pode não atrapalhar muito, né, Aline assim de repente a pessoa não já tá acostumada com Netflix travando com YouTube travando eu não se incomoda tanto entendeu? Que aí você tem as vantagens que você falou a pessoa não precisa baixar um aplicativo, né É muito mais acessível nesse nesse sentido, entendeu?

[41:40] É bom e às vezes pela própria para baixar um aplicativo, ele tem que ser muito funcional, né? Para você ter no seu celular.

Investigadora: Pode explicar sobre a GAB em detalhes? E como estas tecnologias foram usadas nesse personagem?

Priscila Ariel:

[42:03] Bom aí tem a Gabi, né? Como eu te falei que ela é mais jovial e ai eu tô remoto aqui. Deixa eu ver se eu tenho ela aqui, espera aí eu tenho uma aqui.

[42:17] Meu nome é Gabi e quero falar sobre futebol de casa a partir do dia 14 de Maio na quinta serão seis na verdade. É a Gabi é ao contrário HB. Aida mais jovem tá? A gente modificou o personagem botou ela mais jovial, mas é o mesmo é o mesmo procedimento a diferença, Aline é que tá vendo essa tela aqui, ela tem input de texto. Então assim o jornalismo aqui da Globo na época se assustou disse tudo automático e eles não terem controle sobre o que está sendo gerado ali, né? Tá sendo dito e aí a gente botou a Gabi como esse input de texto, tá? Que aí você pode colocar uma imagem um vídeo aqui do lado atrelado essa frase que ela falou então por exemplo se as pessoas para carregar aqui, mas por exemplo.

[43:17] Você tem ali vários Campos de texto. Você pode adicionar vários várias frases, vamos e para cada frase, você pode exibir um tipo de conteúdo uma imagem um vídeo então quando ela sair da frase um e para para falar a frase dois você já muda a

imagem que ela tá exibindo na TV, entendeu? Já já pode adaptar de acordo com o que você desejar.

[43:43] É tá online, mas assim teria tá vendo esse primeiro campo aqui teria vários, tá? Você pode adicionar aqui em cima o quantas frases você quer que ela fale e aí você aqui tem os acessórios ou você consegue por enquanto ela tem um fone um óculos tá que ela não foi muito usada, não. Foi só um pedacinho da pandemia que usaram ela você pode personalizar blusa pode carregar uma imagem aqui e exibir uma blusa uma estampa diferente pode escolher a cor da blusa. Enfim uma customizaçãozinha, né? Que a gente dependendo do quem usasse a Gabi ficaria com a estampa da no grupo que tá usando a Gabi, né de esporte TV ou com a estampa de ela, ela foi usada muito internamente para divulgar notícias, então a Gabi já foi usada assim então por exemplo. Ah tá divulgando a notícia de academia de tecnologia que são treinamentos internos, aí ela vinha de amarelinho como escrito academia de Tecnologia.

[44:43] Aí é uma customização assim.

[44:49] Aqui é legal. Isso é legal de mostrar mesmo. Porque daí aqui a pessoa que tá no workshop pode cansar. [45:04] Ah, eu acho que falando de chefe cheque seria interessante.

[45:20] É verdade, deixa eu tentar abrir local que eu tô ficando irritada com ela. Esse vídeo demorando da minha força. Aí tem que fazer o download. Mas eu vou vou depois eu te mostro os vídeos, mas na verdade vamos continuar perdendo tempo aqui falando de conceito que a gente pode usar também aí esse link eu te mandei até no comentário lá desse eh, quem te publicou em 2020. Acho que você conheceu a Gabi até 2019 assim, mas depois a gente conseguiu lançar um paperzinho e descer com alguns avanços de tecnologia na nuvem e no IBC, quando a gente apresentou ele tá ela tava local com algumas coisas de processando na nuvem. Aqui seria a plataforma toda na nuvem.

Investigadora: E como percebe o uso de deepfake?

Priscila Ariel:

[46:32] Aí como eu te falei tem uma diferença de animal um personagem e o Deepfake essa primeira etapa aqui é como os animadores 3D fazem um personagem

virtual 3D eles pegam o que eles aprenderam de anatomia facial, vamos assim comportamento, né da boca e anima um personagem certo então quando a gente faz o Lipe sim é baseado nesses conceitos de anatomia, né? O quanto você tem que deformar a boca, enfim o rosto para simular aqueles fonemas que tá sendo falado então por exemplo se eu quero que a pai da falhar eu vou ter que abrir a boca dela assim esticar que enfim você movimenta ali a malha três dedos usuário certo aquele aquela superfície

[47:29] no Deepfake você eh não tem esse trabalho a máquina aprende como visualmente se comporta a boca aos olhos e ela reproduz aquilo de forma muito mais fácil. Vamos dizer assim, tá? Você não precisa de uma pessoa trabalhando nesses movimentos. Você pode ter esse movimento automático feito pela máquina e como se você não poupasse várias várias etapas ali do processo do 3D ou 3D tem esse trabalho Inicial como eu te falei que você vai deixar em 3D programado. Como é o a como eu ver e depois você só vai dar comandos para que você executado uma vez produzido você não se preocupa mais com isso, você pode usar esse personagem 3D igual a gente faz com a Gabi com a égua. Vai usando quantas vezes quiser.

[48:27] O deep fake é ele vai ter que passar por esse processamento todas as vezes que você quiser gerar um vídeo novo.

[48:37] Então assim ele já tá treinado vai ser um processamento muito mais rápido, tá? Porque o treino demora a gente em torno de duas semanas para você para máquina aprender a falar com uma Aline tá? Vai botar vários vídeos da Aline aí o Deep Learning vai aprender como Aline fala se expressa e a vantagem do de fake é que quando eh o vídeo novo que você está gerando tem um sorriso não vai ser um sorriso genérico vai ser o sorriso da Aline, porque às vezes a gente tem umas expressões únicas, né, meu sorriso, por exemplo é mais para um lado do que para o outro, né? Levanta mais uma coisa do que a outra e cada um tem essas particularidades a vantagem de fake é que ele traz todas essas particularidades da pessoa. Por isso também que convence mais de ser algo real. Tá do que o personagem 3D o personagem três dele é um pouco mais genérico, vamos dizer assim.

[49:37] Não tem tipo, se eu levantar a sobrancelha vai levantar um pouco igual não tem tanta imperfeição tanta assimetria. Vamos dizer assim facial. E aí eu Deep fake ele consegue trazer mais Realismo. Por isso também ele aprende como cada indivíduo

se comporta como se expressa e aquilo reflete também nos vídeos que eles geram. E aí você acaba acreditando que aquilo é real, entendeu? Tem tem isso, né? Tem essas particularidades. E aí é

[50:12] Conseguiu colocar não, mas ele aqui tem um videozinho. Deixa eu tirar o áudio em tempo real, ó. Isso aqui já tá em tempo real. A gente já consegue usar uma é código aberto ali, né? Por isso que a gente tá vendo tanto de fake isso aqui tá no github, você consegue baixar essa ferramenta? Se você tiver uma máquina potente uma placa de vídeo boa, você baixa essa ferramenta que é o Deep Face Lab é o nome da ferramenta que tá no github você consegue é botar nessa ferramenta é a treinar com outra que tem duas ferramentas uma de treino e uma de você executar aquilo ao vivo.

[50:59] Então na de treino. Você consegue colocar vídeos seguir passo a passo ali do que tá descrito no github de como você fazer o treinamento. Aí você coloca vários vídeos da pessoa que você quer treinar vários vídeos também da pessoa que você quer substituir. E aí você consegue fazer essa relação, né? Quando essa pessoa abre a boca, eu quero exibir o ar dessa daqui e aí você substitui o rosto fazendo os mesmos movimentos que a que é anterior faz, entendeu? E aí você consegue gerar isso qualquer um já era isso. É isso que eu tô te falando, por isso que a gente tá tendo tanto de fake em relação a vídeos a vezes é tudo tá porque tá tudo aberto.

[51:46] Você tá até tem tecnologia disponível no github, se você tiver uma máquina potente para processar isso você consegue baixar e gerar seus vídeos no rosto das pessoas, né? E assim a gente tenta usar de forma consciente, mas tem gente que não usa né? E aí gera essa desinformação e tudo isso então a gente tá você tá se preocupando, né em ter uma checagem de fonte e tal, mas eu acho que isso vai além. Sabe tem que ter mais rígidas, para que isso fique disponível só para pessoas que se comprometam em produzir conteúdo de forma legal, entendeu? Tem muita mas coisas envolvidas.

Investigadora: Isso implica nas questões éticas sobre uso de deepfake?

Priscila Ariel:

[52:36] Porque assim se essas tecnologias ficarem disponíveis que está no Brasil, isso é é muito importante de ser aprovado, porque a responsabilização inclusive das empresas, né? Então é justamente isso que você tá falando.

[52:52] Hoje tá ali disponível tá e a liberdade, mas a liberdade de responsabilidade, né? E aí esse que eu tô te falando. Por isso que a gente tá vendo muito assim, porque algo que as pessoas venham Realismo então é fácil você produzir e as outras que assistem acreditarem naquilo que você produziu porque ele já tá gerando uma qualidade de de substituição de rostos muito boa, assim que você se você gerar um vídeo de alta resolução, você vai ver a imperfeições ali identificar que é um que é onde perfeito, né? Mas se você gerar vídeo de internet de WhatsApp que você diminui a qualidade cara, fica muito muito assim aceitável, sabe da pessoa acreditar que aquilo é real porque você tá botando você tá usando a baixa qualidade de

[53:50] De disponibilização de conteúdo e a pessoa vendo uma telinha de celular, ela acredita sabe que é o Bolsonaro que é o Lula que é enfim qualquer pessoa ali falando o que foi gerado de forma sintética, entendeu? E aí isso aqui que eu ia te mostrar. Deixa eu ver deixa eu deixar processando aqui para ele abrir ó, tá com áudio melhor contou.

[54:58] Agora tá agora sim, vou dar um play só ver se você escuta ela conectei mas eu vou te mandar essa apresentação em resumo, ela tá falando em vários idiomas é a mesma atriz de geração de voz automática e substituição de rosto. Eu gostaria é só substituição de rosto nesse caso nesse caso que é outra pessoa falando o idioma correto. Mas você usa a mesma atriz original do vídeo, entendeu? Então você faz consegue fazer uma correlação de uma voz usando outro rosto.

[55:46] Tá então a sentir ela ela usa isso para você fazer transcrição tradução de idiomas, né? Você consegue pegar o mesmo conteúdo, por exemplo gravar um comercial, né com alguma ator esse ator não precisa aprender vários idiomas, né para você exibir aquele comercial em outros países, você pode pegar o mesmo vídeo desse ator fazer um treinamento de de pillary, né com ele para poder gerar para máquina aprender como eles comportam como ele fala e aí você encurta outros áudios ali para ele falar e esse áudio é alguém falando certo em outro idioma, entendeu? Então.

[56:29] Falando a mesma coisa em outro idioma é uma forma de você fazer uma dublagem perfeita. Tem um cinema que às vezes a gente vê aquela dublagem estranha, é porque eu a dublagem tradicional tá fazendo em cima de fonemas de outro idioma, né? Então tipo, ele tá tentando bom em cima do da Minnie que em inglês aí

não casa, né, mas não casa, entendeu? Aqui não além de mudar o idioma também muda o fonema esse essa esse lip syn que essa sincronização labial também é alterada isso aqui facilita é geração de conteúdo para vários países assim tanto no cinema quanto na propaganda marketing. E como tá acessível, vamos dizer assim cada vez mais acessível. Essas tecnologias daqui a pouco o YouTube, você vai conseguir gerar o seu conteúdo, né de YouTube ali.

[57:27] Usar ele reutilizar para lançar ele em outras plataformas com outros idiomas, né principais ali então enfim isso tá popularizando e cada vez mais teremos essa tradução de idioma de forma mais natural usando um único vídeo baseado.

[57:49] Aí o que a cientista faz disponibiliza além disso aí que eu te mostrei é o que você ter vídeos corporativos, né de uma pessoa falando um conteúdo que você quiser uma aula uma notícia, talvez poderia ser um formato assim como um propósito também para o seu assistente.

[58:13] Entendeu de você ter ali alguém na frente gerando texto atrás trazendo imagens relacionados aquilo que você perguntou entendeu? Aí outra forma aqui é de você ver pessoas mortas voltando a vida, né? Aqui é o Salvador Dali e ele interage com quem viver com museu é em São Peters, se eu não me engano. Esse museu. E aí ele tira selfie com as pessoas, né? Fala da história dele é uma forma legal de usar o defeito, entendeu? Tipo você trazer da vida aí essas pessoas que morreram. Só que tem uma discussão ética também isso né de você gerar conteúdo novo e a pessoa tá morta. E aí?

[59:09] Tem gente que não gosta, né? Ah, mas o Salvador Dali não falaria isso, você está manipulando a imagem de Salvador Dali. Enfim, tem muitos problemas relacionados, Aline tu tá com horário eu tô Seguindo aqui, mas nem sei se você tá não quero te atrapalhar também, mas aqui Você já viu provavelmente que é aquele site de chinês, né? Que usa fake também fica 24 horas ali rodando notícias aí é o que eu te falei, né? Fizeram um contrato com esse repórter aqui para escanear e usar a imagem dele ali, né? Mas você se repórter for ali na rua aprontar alguma coisa.

Investigadora: Pode destacar outras soluções feitas ainda na Globo com uso destas tecnologias?

Priscila Ariel:

[59:57] Tem vantagens tá você consegue rejuvenescer pessoas usando deepfake, então aqui é um exemplo. Você é outro exemplo para rejuvenescer, né? Aí aplica ali um filtrozinho para não mostrar as imperfeições.

[01:00:33] Ah agora eu preciso de uma cena que a pessoa através disso já tá vindo, cara. A gente a gente tava tentando emplacar isso tá vindo agora isso aqui eu vou te dar um pause aqui no ABS, pera aí tem a substituição de roxo de dublê rejuvenescimento, né? Como eu te falei um custo menor. Porque vão ter menos pessoas dedicadas a isso e vai produzir mais rápido tá então antes de demorava mesmo para substituir um dublê um rosto numa cena agora você depois que treinou essas duas semanas de Treinamento que eu te falei de rosto. Você já consegue gerar de forma simples sabe em horas dependendo do tamanho da cena, né? Você gera ali em alguns minutos ou horas se for uma cena muito grande.

[01:01:21] Entendeu? Então você economiza muito tempo e reaproveita material, né? Você usa aquele treinamento se o ator não mudar você vai usando aquele treinamento ali para as próximas cenas. Aí também tem a possibilidade de juntar os dois tá Aline que é o que a gente tem feito agora também se eu pegar a Aida e jogar o DIP fake. Deixa eu só tirar o áudio aqui para poder falar contigo ó, a gente jogo de fake da Maju em cima e traz mais Realismo também então ali você a gente já fez projetos aqui que eu vou te mostrar que os outros dois tá a gente produziu em 3D e trouxe mais Realismo pro rosto que é onde apresenta onde o três de pé que é mais no rosto, sabe que entrega que não é humano e aí quando a gente joga um defeito traz mais Realismo mais naturalidade convence que alguém mais assim no rosto, né? É bom tem aqui.

[01:02:19] Recém labial que é o que eu te mostrei igual da cientista, né de você, por exemplo, ela tá falando uma coisa, mas eu posso treinar ela para falar o que o Adnet tá falando aqui do lado, entendeu?

Eh, sei que você não tá escutando aí que eu não consigo compartilhar áudio aqui, mas você consegue mudar o que é falado tá aqui a gente também usa o de fake agora para gerar novos conteúdos 3D.

Porque a gente percebeu uma linha de pesquisa aqui que você consegue trazer mais Realismo por 3D. Se isso for feito a partir de um defeito e trazer movimentos mais realistas. Aí a gente tá treinando também tá botando em cima de uma pessoa.

[01:03:08] Em vez de você capturar com o iPhone de movimentos ali e passar você treina na parte de feio que traz um pouco mais Realismo para esse 3D.

[01:03:23] Ainda é uma linha de pesquisa, mas a gente tá vendo. Que que isso pode trazer benefícios tá aí aqui tem exemplos, né que a gente foi fazendo botando Cauã na cara das gêmeas de novelas. E aí divulgando uma das novelas aqui.

[01:03:41] Era Rute, Raquel. Essa novela aí de era para botar no Instagram e divulgar a novela teve a Tatá Werneck também agora foi ano passado que a gente fez trocava o rosto com o convidado.

[01:03:55] E aí ela com a cara e vice-versa, né ele com a cara dela e foi vários quadros, ela usou com Tony Ramos ou com várias pessoas aqui é o jeito que a gente grava tá Aline tá tão fácil de usar o defeito hoje em dia que a gente pega o ator, ele assinou o contrato lá dizendo que não pode usar, né? A gente pega o ator. Bota ele sentado com luzes, né em volta uma câmera boa na frente dele câmera normal, pode ser até um iPhone, né? Que também tem uma qualidade boa de vídeo e aí você pede para ele falar umas frases que tem já uns fonemas específicos ali que é importante ter que aprender e você tem os movimentos de cabeça dele falando naquela frase. Não sei o que você grava aquilo.

[01:04:42] 10 minutinhos da pessoa falando coisas movimentando você muda a luz apaga um lado apaga o outro para poder ter vários ângulos de iluminação e casar com os próximos vídeos e aí você joga esse videozinho que já já que você gravou do modo ideal podia ter aqui ó. E você já consegue treinar essas duas semanas com movimentos e você usa já no ao vivo também a Tatá Werneck, eu vou dar play aqui, você vai ver, ó.

[01:05:15] Que a gente fez essa etapa do treinamento e aqui ó, tá sendo usado ao vivo no programa. A Tatá Werneck num monitor você precisou treinar os convidados isso. E aí a gente levou esse processamento numa máquina lá do estúdio, ele te levou a máquina também, né? Porque eles não tinha. E aí a gente processou isso em tempo real lá então o vídeo que chegava na máquina, né? Tanto da Tatá e do convidado a gente processava em tempo real trocando os gostos e o resultado da gente jogava numa tela lá do do palco, né? E eles conseguiam assistir em tempo real. O Tony Ramos ficou muito bom, né? Olha ele com a cara dela, eu vou te mandar que eu te mandar os links também que eu tenho que te mandar essas apresentações enfim.

[01:06:12] E aí assim isso já tá sendo usado o JN usou também a gente ele fez uma série de Constituição explicando a constituição e tal e um deles era desinformação um dos temas de fake mesmo. E aí eles pediram para gente fazer um exemplo de fake para eles mostrarem e aí aqui é o bastão repórter bastante como putinha aí ele fala que no vídeo aqui ó, ninguém acredita que o Putin vai falar português de uma hora para outra então não confie no que vocês vêm aí é ele com a cara do pudim. E aí depois a gente desmente isso, né? É verdade?

[01:07:12] Isto e e virando bastante de novo, mas é foi para alertar, né? Usou mas usou realmente de forma negativa, mas existe o lado negativo, né isso porém e acho que impacta muito forte o jornalismo, né? É o jornalismo impactado muito de forma negativa com isso, né o entretenimento não né? Pode usar como igual a Tatá, né? Como entretenimento mesmo não tem muito risco, mas realmente jornalismo é tem o pacto negativo. E aí aqui é um exemplo do gênero Escobar que a gente também juntou os dois tá 3D com com o Real. Então é o gênio é 3D que a gente usou na Copa agora da da última. Copa a gente treinou o Escobar caracterizado com a barbinha do gênero. Não precisa pintar ele de azul, nem de verde, tá? Mas a gente jogou o azul depois mas a gente tinha um ator ali fazendo os movimentos do

[01:08:11] Mais de 3D e por cima jogava eh a voz de Escobar e o o o Deep fake para trazer mais Realismo então aqui a gente tem o 3D no nessa primeira parte aqui que já ficou legal também tá? Mas quando a gente joga o de fake fica muito mais realista, entendeu? E aí a gente consegue trazer um convencimento maior das pessoas é isso que eu tô te falando, se a gente jogar só o 3D as pessoas poderiam achar um meio estranho Escobar falando Escobar né? Se comunicando ali, mas quando a gente joga de fake essa estranheza diminui muito sabe?

[01:08:54] Enfim é mais um ponto de vista de convencimento daquele conteúdo sabe de trazer mais naturalidade para aquele conteúdo e aí eu te mostrei alguns exemplos, eu tinha separado outros também aqui. Ó, esse cara aqui, não sei se ele vai me ajudar agora porque tá remoto, mas tem um link aqui, esse cara é um outro exemplo de uso de jornalista. Não sei se você já viu também aqui eu já te mostrei deixa ver se tem mais alguma coisa nesses aqui.

[01:09:30] Tô vendo só tem mais algum conteúdo diferente, a gente já uso para caramba no Google assim, mas só para entretenimento aqui foi na Fátima. Onde o tentava avançar a gente fez um efeito do Daniel falando?

Coisa que não falou né? Aí a Fátima falou assim. Ah Daniel, tem uma notícia sua legal que você falou isso aí era o Daniel falando que ia fazer parceria com a Ludmila e que iria para o funk sabe? Aí ele ficou olhando assim, tipo, eu nunca falei isso, né aí e aí também foi um tipo de alerta no programa, né? Para você não acreditar em todas as notícias. Enfim, tá remoto, ele não vai me ajudar agora a gente já usou um programa que não deu certo que foi o caso da kalima, é?

[01:10:25] Mas era para fazer de fake de um quadro também dentro do programa para fazer de fake dos Big Brothers trocar o rosto dos Big Brother. Aqui é uma abertura do fora de hora, onde a gente substitui o rosto de

Coisas jornalistas, eh, deixa eu tirar o áudio para mim poder falar junto vídeos de jornalismo famosos assim que imprevistos acontecer tipo placa caindo atrás da cabeça jornalista. Aquele caso também do menininho que entra atrás enquanto o pai tá falando ali na política, a gente foi substituindo pelos rostos dos comediantes, né desse quadro aqui do Fórum de hora e aí enfim já usamos algumas coisas, né de fake aqui na casa a maioria para entretenimento.

[01:11:17] Mas eu acho que no geral é isso. A linha, deixa eu ver se tem alguma coisa local aqui que eu possa te mostrar.

[01:11:27] Agora abriu ó. Agora abriu o local. Mas é isso sabe vídeos de pessoas que viraram meme né na internet e a gente substitui a casa do menininho que entrou Galvão botou a cara do outro. Enfim, mas o que eu queria te mostrar mesmo, né? Era essa parte tecnológica, né dizer que tem essas diferenças entre tecnologias e que Deep fake é só para vídeo quando você diz esses vídeos sintéticos aí é tipo fake, se você quiser falar, por exemplo de voz sintética, aí você pode dizer da tecnologia por trás, né que tá em todas né? Que é o depilando.

[01:12:14] Que esse aprendizado profundo de máquina, entendeu? E aí esses conceitos vão ser importantes para você também. Quando falar deles, né você saber essas particularidades, né? E não falar algo assim trocado, né? Na hora da apresentação, entendeu?

Appendix L

Mapa de Empatia

Grupo: 1

Personagem: Ti João J

1. O que é que pensa e sente:
 - O que é que dá sentido à vida?
 - O que o preocupa?
 - O que os faz felizes?
 - A comunhão de sangue e de coração, fazer felizes os outros.

2. O que é que veem?
 - O que vê no seu entorno imediato?
 - Falta de carinho.
 - Desinteresse pelo próximo.
 - Egoísmo.
 - Preconceito.

3. O que é que eles dizem e fazem?
 - O que é que o ouvimos dizer?
 - O que podemos imaginá-lo a dizer?
 - Que comportamento temos observado?
 - O que precisa fazer de diferente?
 - Palavras de carinho.
 - Gargalhadas.
 - Boa disposição.
 - Ditos antigos.
 - Ouvir e estar atento ao outro.
 - Ajudar quem necessita.

4. O que é que eles ouvem?

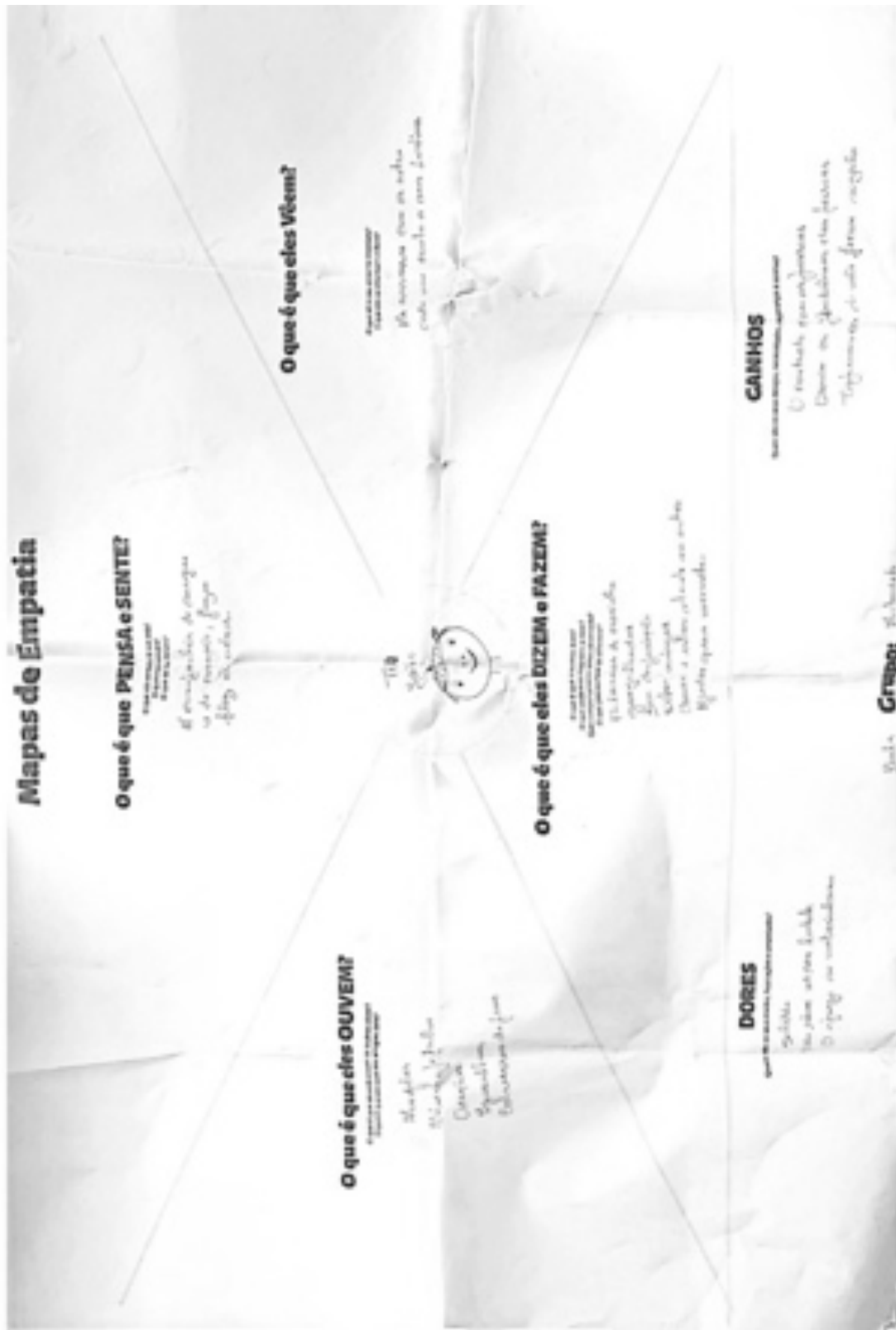
- O que é que está a ouvir os outros dizerem?
- O que é que ele ouve dos amigos dele?
 - Anedotas.
 - Música popular.
 - Orações.
 - Agricultura.
 - Conversas da fama.

5. Dores

- Quais são os seus medos, frustrações e ansiedades?
 - Solidão.
 - Não estar no seu habitat.
 - O apego ao materialismo.

6. Ganhos

- Quais são os seus desejos, necessidades, esperanças e sonhos?
 - O contacto com as pessoas.
 - Ouvir os problemas das pessoas.
 - Esperança de não ficar sozinho.



Grupo: 2

Personagem: Coruja – O Sabe tudo

7. O que é que pensa e sente:

- O que é que dá sentido à vida?
 - A família.
 - Ser feliz.
 - Ter saúde.
 - Convívio.
 - União.
- O que o preocupa?
 - Desigualdade.
 - Solidão.
 - Abandono.
 - Falta de amor.
- O que os faz felizes?
 - Penso que a empatia vai ser a solução à resolução da qualidade de vida dos nossos velhinhos.
 - Tentar ser feliz.
 - Ter pessoas a minha volta.

8. O que é que veem?

- O que vê no seu entorno imediato?
 - Falta de carinho.
 - Desinteresse pelo próximo.
 - Egoísmo.
 - Preconceito.

9. O que é que eles dizem e fazem?

- O que é que o ouvimos dizer?
- O que podemos imaginá-lo a dizer?
- Que comportamento temos observado?
- O que precisa fazer de diferente?
 - Estuda muito .
 - Vê tudo e ouve.

- Partilha saberes.
- Especula pela certeza
- Amor e carinho

10. O que é que eles ouvem?

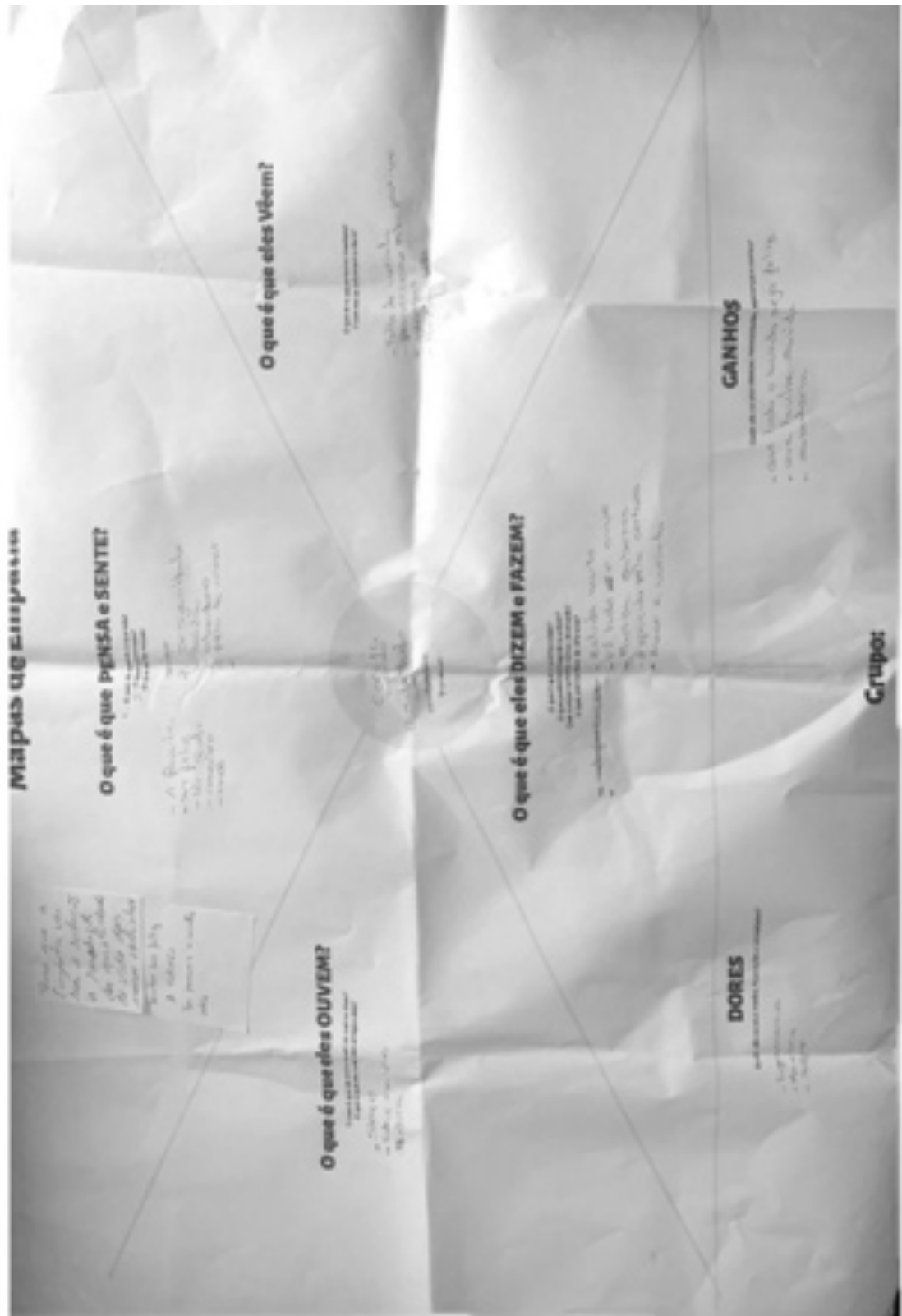
- O que é que está a ouvir os outros dizerem?
- O que é que ele ouve dos amigos dele?
 - Silêncio.
 - Fazer em vão.
 - Mentira.

11. Dores

- Quais são os seus medos, frustrações e ansiedades?
 - Ignorância.
 - Mentira.
 - Sofrer.

12. Ganhos

- Quais são os seus desejos, necessidades, esperanças e sonhos?
 - Que todo mundo seja feliz.
 - Que tenha saúde.
 - Sabedoria.



Grupo: 3

Personagem: Voz idêntica à Gata

13. O que é que pensa e sente:

- O que é que dá sentido à vida?
 - Clareza da sua informação.
- O que o preocupa?
 - Veracidade da informação.
- O que os faz felizes?
 - Que os outros a sintam como verdadeira.

14. O que é que veem?

- Acontecimentos e factos que precisam esclarecimento
 - Base de dados fidedigna e ou vídeo na Internet ou presencialmente (cruzar informação).

15. O que é que eles dizem e fazem?

- O que é que o ouvimos dizer?
- O que podemos imaginá-lo a dizer?
- Que comportamento temos observado?
- O que precisa fazer de diferente?
 - Intro clara e imparcial.
 - Pensamento neutro.
 - Comportamento sério.

16. O que é que eles ouvem?

- O que é que está a ouvir os outros dizerem?
- O que é que ele ouve dos amigos dele?
 - Outros: dúvidas, falsas informações.
 - Amigos: preocupações.

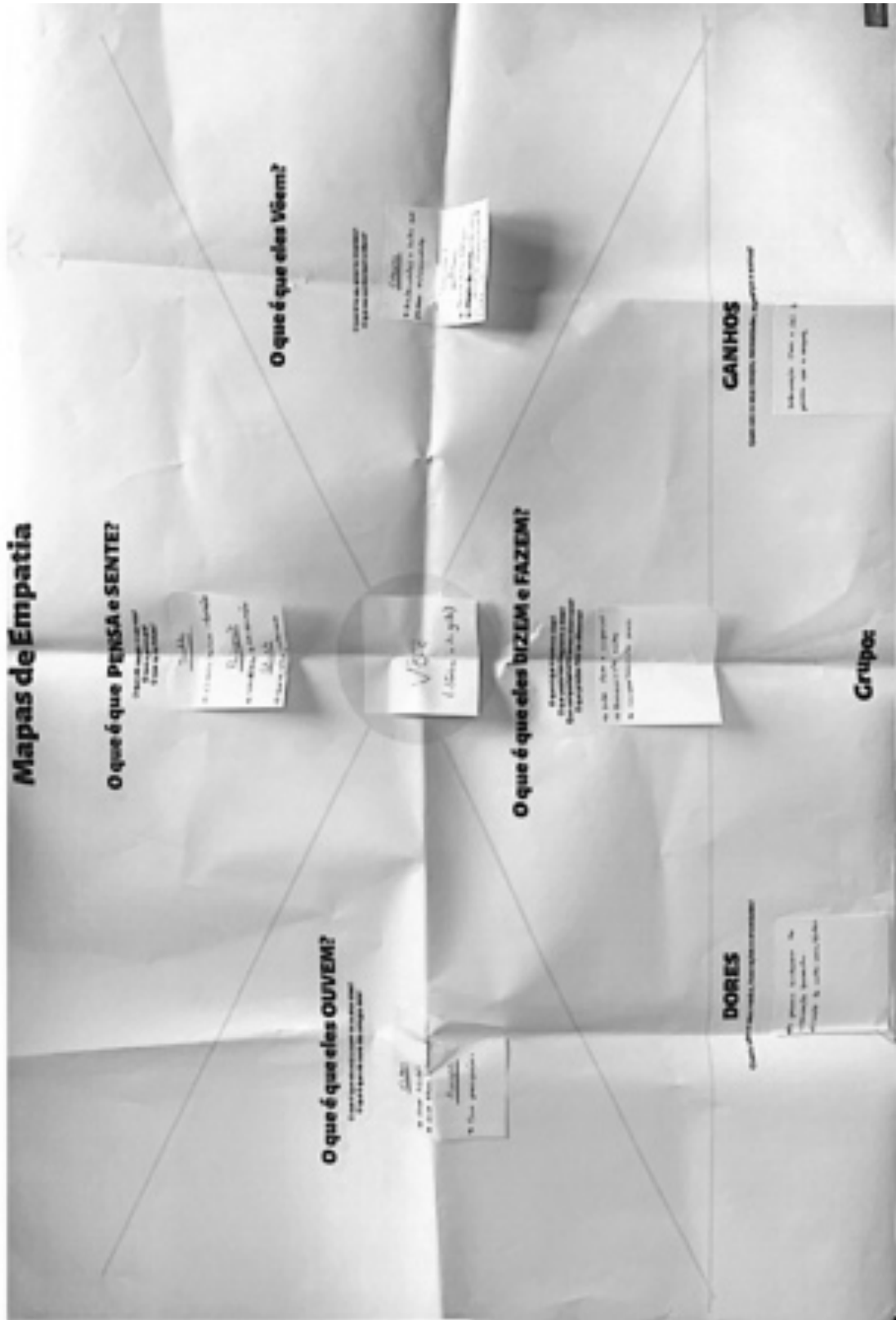
17. Dores

- Quais são os seus medos, frustrações e ansiedades?
 - As pessoas duvidarem da informação fornecida.
 - Criação de ainda mais dúvidas.

18. Ganhos

- Quais são os seus desejos, necessidades, esperanças e sonhos?

- Informação clara e útil à pessoa que requer.



Appendix M

Transcrição de áudio

Grupo 1

PQ6 faz a apresentação do grupo:

[00:00 - 00:52] É assim, nós escolhemos uma personagem inspirada no Rui Costa (Rui Costa, que está na sala fotografando: - falsas, falsas), inspirada em locutores de rádio como Rui Costa e o ti João de Bragança. E conseguem (PQ10: - Que é mais velhinha) chegar mais longe em termos de comunicação e falar com eles. Tem uma diferença em relação às personagens deste projeto: Têm a capacidade ouvir. Essa poderia ser uma dica para deixar: que tivesse a capacidade de escutar os problemas das pessoas. Porque ao longo das coisas todas que fomos acrescentando a escuta foi um aspecto fundamental.

[00:53 - 01:18] é então, tínhamos então o que eles ouvem? Por que as pessoas se identificam com ele: anedotas, música popular, as orações que eles tanto gostam de ouvir, agricultura e conversas do povo, ou seja, temas mais indicados ou mais orientados para as pessoas do dia a dia.

[01:19 - 01:35] O que é que ele pensa, e sente? A companhia, as palavras do coração, e o carinho, e tenta sempre fazer felizes os outros através de mais uma escuta.

[01:36 - 02:24] O que é que eles dizem que fazem? Tentamos focar mais no que é que ele teria que dizer e como é que ele teria que falar. Seria então, com palavras de carinho, com muitas gargalhadas, como sempre bem-disposto; que isto em Avatar é sempre mais fácil, e com ritos antigos e populares; mais uma vez viemos buscar os temas supridos anteriormente: ouvir e estar atento ao outro e ajudar quem necessita. A tentativa de que é uma coisa que este projeto já tem de dar resposta às necessidades e de responder às perguntas e mais uma vez tentar que houvesse uma forma em que pudesse escutar. Não ser só da pergunta básica.

[02:25 - 02:52] Então, o que é que eles veem? Então teremos aqui as conversas com os outros e em que cada um conta a sua história. Mais uma vez a personagem teria que ter estes aspectos em conta da relação de uns com os outros e de criar esse espaço de escuta. Que fosse maior do que uma pergunta que fazemos. Se há alguém que responde às nossas perguntas já é um espaço de escuta, não é? Mas que fosse um bocadinho mais além do que isso.

[02:53 - 03:03] Esta é a nossa sugestão e o nosso bonequinho seria inspirado no tio João de Bragança, que com os traços do Rui Costa.

Grupo 2

PQ3 faz a apresentação do grupo:

[00:00 - 00:24] Portanto, o principal é família. Temos alguém da família à nossa volta para nos auxiliar e ser feliz e ter saúde, o convívio, a união, a culpa, a desigualdade, a solidão, o abandono, a falta de amor.

[00:24 - 00:39] E agora aqui, o que é que eles ouvem? O silêncio. Muitas vezes não são corretos conosco e ficam em silêncio.

[00:39 - 00:53] falar em vão, às vezes falamos ou ouvimos a opinião dos nossos amigos, supostos amigos, e não nos ouvem nada. A mentira. Muitas vezes é mentira.

[00:54 - 01:08] O que é que eles veem? A falta de carinho, o desinteresse pelo próximo, o egoísmo, a persuasão.

[01:09 - 01:24] O que é que eles dizem e fazem? A coruja. O que é que eles nos permitem? Que ela estuda muito, vê tudo e ouve, partilha os saberes e especula pela certeza, amor e carinho.

[01:24 - 01:47] E as dores. As dores: É a mentira, o sofrer.

[01:48 - 01:49] E os ganhos, como todo o mundo seja sempre é feliz, que tenham a saúde e sabedoria.

Grupo 3

PQ1 faz a apresentação do grupo:

[00:00 - 00:10] Nós não escolhemos nenhuma personagem, porque entre cada um de nós escolhemos personagens diferentes... Mas aquilo que mais recomendamos é que o tom de voz ou o áudio que deve ser acompanhado sempre que seja texto ou de personagem e que nos crie alguma proximidade. Por isso, o tom de voz que nós escolhemos era ali da gata Melinda, que é aquela que às tantas dificuldades pode ajudar-te nessa passagem de informação. Por isso, nós escolhemos uma voz com o áudio, para não teres assim esses sentimentos como se tivesse uma pessoa.

[00:33 - 00:46] Mas poderá dar sentido a esta voz, a mesma clareza da sua informação. A preocupação é que esta clareza seja útil a quem for procurar no buscador a dúvida que possa colocar. Portanto, houve muitas dúvidas, houve muitas

informações falsas, quer dos outros, quer dos amigos, também que vão procurando ter essas preocupações. Mas todos nós... Nós fazemos a mesma coisa. Ouvimos aqui e vamos para ali.

[01:02 - 01:14] E depois, andamos aqui nos computadores buscar essa informação. Portanto, essa voz na informação, aquilo que vai querer também ele, procurar ver se esses contos vividos que são verdade, vão cruzar essa informação toda, que nos vai a nós também dar essa verificação da veracidade desses questionamentos.

[01:27 - 01:38] Portanto, dores e ganhos. Dores não têm muitas, a dor dela é se eventualmente não passava a informação e não cria a empatia com quem procurar essa mensagem. E o ganho será sempre para quem criou a aplicação. Vai ajudar todos aqueles que vão usar a aplicação.

[01:39 - 01:51] Agora, creio que a imagem não cria tanta diferença. A voz e o tom de voz, ou o áudio, é que se calhar cria mais proximidade e garante alguma veracidade daquilo que nós estamos a ouvir.

[01:51 - 01:57] Ou esse sentido de nos ligarmos a essa mensagem, não sei. Mas é aquilo que nós sentimos. Muito obrigado.

Appendix N

Focus Group script questions.

1. O que é que mais gostaste ao interagir?
2. Houve alguma coisa que achaste difícil durante a interação?
3. Sentiste que respondeu às tuas necessidades de informação sobre saúde e envelhecimento? Por quê?
4. Qual foi a solução preferida e por quê?
5. Prefere receber a solução imediatamente à pergunta ou não se incomoda que demore uma semana?
6. Havia alguma informação importante que esperavas encontrar, mas que não estava disponível?
7. Achas que seria útil para outras pessoas da tua faixa etária? Porquê?
8. Como te sentirias ao recomendar o uso deste chatbot a outras pessoas na tua comunidade?
9. Acreditas que o chatbot pode ajudar a reduzir a disseminação de desinformação sobre saúde entre mulheres idosas?
10. Este tipo de assistência automatizada é algo novo para ti?
11. O chatbot captou eficazmente a tua atenção? Porquê?
12. Que melhorias poderiam ser feitas para aumentar a tua confiança no chatbot?
13. Sentiste-te envolvido(a) durante a interação com o chatbot?
14. Que elementos te fariam sentir mais familiarizado(a) com o chatbot?
15. O chatbot de alguma forma provocou medo ou insegurança? Por favor, explica.

Appendix O

Focus Group transcription.

[00:00 - 00:07]

- Investigadora: Alguém queria comentar, fazer um comentário, o que achou, qual foi a solução que achou mais interessante?

[00:10 - 00:24]

- PQ2: Eu gostei da Dina Aguiar, eu sabia que deitar cedo podia melhorar na tensão, mas não sabia que não influenciava deitar cedo e também porque é falso, mas gostei de entender.

- Investigadora: A informação foi interessante, então? Querem comentar alguma mais alguma coisa sobre isso?

[00:37 - 00:47]

- PQ9: A dica sobre o AVC também gostei de saber. Fiquei até preocupada, fiquei assim um bocadinho mesmo.

- Investigadora: Vocês acham que a informação é importante, assim, até para ter um embasamento?

- PQ4: Sim, para os que tiverem um pouco de juízo foi.

[00:47 - 00:59]

- PQ8: Sobre tudo, muito elucidativa, compreensiva, principalmente a solução em áudio e texto. Para mim foi top.

- Investigadora: Ah é? Preferiu aquela e não as em que as pessoas falavam?

[00:59 - 01:10]

- PQ8: Sim, o texto e a voz. Acho que foi mais direto.

- PQ5: Eu acho que vai depender muito da pessoa que está por trás do ecrã. Pessoas que estão habituadas a muito texto, como eu, por exemplo, eu que estou habituada ao texto, achei o texto com o áudio mais abrangente. Os personagens não me dizem muito. Há algumas que gosto, mas não me dizem muito. Mesmo a solução do Polígrafo não me diz muita coisa. Agora o texto e o áudio sim. Mas se calhar pessoas de uma geração mais acima, se calhar tem uma Dina Aguiar que já é conhecida, as pessoas estão familiarizadas, já dá outro conforto. Acho que vai depender muito da pessoa que está por trás do ecrã.

[01:38 - 01:48]

- PQ6: Eu acho que a maior parte das pessoas gostam do real, gostam da pessoa, da pessoa no vídeo. Não é do animal, não é do virtual. Os mais novos são mais pro virtual, como aquele 3D. Mas nós não.

- PQ3: Ah, mas eu gosto.

[02:09 - 02:20]

- Investigadora: A primeira personagem que passou para vocês, a Sônia, gostaram?

- PQ8: Para mim, eu gostei, o que importa é a mensagem, não interessa se com gato ou outra coisa. É quase como se estivéssemos a conversar com alguém. Embora esteja, no caso, eram os gatos.

[02:39 - 02:49]

- Investigadora: E quanto ao humor, o que acharam?

[03:02 - 03:12]

[00:59 - 01:10]

- PQ8: Para mim mais do que o personagem foi o tipo de tom de voz e a linguagem utilizada. Acaba a ser mais “terra a terra”, quer dizer, mais natural. A personagem, também prefiro uma coisa mais real do que virtual. Mas os bonecos, para mim, prefiro mais os gatos a as outras personagens virtuais. Dá-me a sensação de mais falso quando é uma personagem 3D a imitar alguma coisa que é real. Mas que não é o caso do gato, eu sei que é uma brincadeira interativa. Mas o que depois me conta muito é a voz e a informação à partida também não apresenta uma informação que me leva a pensar que será menos negativa ou falsa. Já uma imagem virtual. Já fico com mais dúvidas em relação a como possa ser a resposta.

- Investigadora: Seria a da Aida, no caso?

- PQ8: Da Aida, da Sónia... A Dina Aguiar. acaba por contrariar isso porque é uma personagem que nós estamos mais familiarizados. Mas se calhar a Aida, no Brasil, tem outra intensidade à informação que passa. A nós não nos diz muito.

[04:13 - 04:25]

- PQ11: Eu prefiro a solução com texto e áudio. Mas se tivesse de ter uma personagem para mim não seriam os gatos não. Porque o gato do jeito que falou passa uma coisa de extremismo. Eu tenho gato e sou muito fã do gato. Mas acho que nesse sentido não

ficou interessante. Não. Acho que ter uma personagem que brinca, talvez pela seriedade do tema. No meu caso, eu prefiro uma abordagem mais clara e específica. Porque se é um tema de saúde, ter um gato a brincar com isso e a falar sobre isso... não me chama a atenção em nada, pelo contrário, faz-me ficar com um pé atrás... Se fosse para criar um ícone só e ter ali em áudio e texto está bom.

[05:25 - 05:35]

- PQ4: Eu gosto que a informação esteja lá. Eu ando sempre à procura de sites científicos. E que haja a referência.

- PQ3: Eu acho que aquelas pessoas que não têm acesso à internet nem sabem trabalhar na internet. Elas ficam fora dessa solução. A partir dos 60 e 70 anos, muitas pessoas não sabem trabalhar na internet.

[06:36 - 06:47]

- Investigadora: É nesse um ponto que eu queria chegar. Com relação aos cuidadores. Porque a ideia é que, de repente, o cuidador possa ajudar. Então, ele vai ter um acesso a esse chatbot e vai poder mostrar esse conteúdo para a pessoa que ele está cuidando. Então, nesse contexto, acham que seria interessante esse tipo de tecnologia ou de solução?

[06:48 - 07:00]

- PQ10: Eu queira ensinar aos cuidadores a ter acesso a isto. É a minha opinião. Porque a maioria dos cuidadores informais não tem.

[07:00 - 07:12]

- PQ6: Mas há aí uma alteração que tem de ser chamada a atenção. É que os cuidadores muito mais também estão a levar a informação. Porque enquanto as pessoas idosas tinham um nível de formação e os cuidadores tinham o mesmo nível de formação, agora estamos na fase de mudança. Eu, por exemplo, tenho um ensino superior, mas sou igualmente cuidadora dos meus pais.

[07:25 - 07:36]

- PQ2: Eu, por exemplo, não tive formação. Mas eu tive que cuidar do meu marido 86 anos. O meu marido morreu por problema no quadril com os 85 anos. Eu tinha 66. Sou bastante ajudativa. Porque tinha uma senhora que me pediu ajudasse com informação.

[08:11 - 08:24]

- PQ9: Eu acho que o nosso país devia estar mais atento. Estar mais atento.

[08:36 - 08:48]

- Investigadora: Porque o verificador é isso. Ele vai lá e mostra. Está falso por causa desse fato aqui. Os cientistas devem ter se separado. Eles falam. O cientista tal e tal disse. Por isso, isso e isso. Ela ajuda as pessoas a entenderem. Olha, lá no link tal, você vai encontrar a informação que foi verificada. E por isso é falso. Porque às vezes na internet a gente recebe coisas e não sabe entender muito bem porque aquilo é verdade ou falso. A desinformação é isso. Ela vem com uma coisa assim muito, nossa, o leite da vaca puro não causa intolerância. Será? Eu tenho intolerância? E você não sabe como buscar essa informação. Então, o interessante é que tem um fator educativo. Mas eu entendo que às vezes precisa de uma informação ainda antes. Que é como acessar ali e mexer. Que é uma interação digital. Então, até uma coisa para o EduCage trabalhar.

[09:45 - 09:57]

- PQ9: Eu tenho uma amiga, né? Que sempre cuidou sempre. E ela própria me ajudou a mexer com a internet. No caso de mim. Nós que temos que criar estratégias.

[10:08 - 10:20]

- PQ3: Eu trabalhei com um lar de idosas na Suíça. Em Louzanne. No Chateau de Lannes. E lá era o máximo. Porque havia três casas na mesma...as que sabiam tratar delas sozinhas. E tinham televisão. Toda a gente tinha televisão. Toda a gente tinha a vontade própria. E no outro setor tínhamos as pessoas acamadas. E que tinham problemas. Poucos problemas. E no outro setor era problema psicológico. Eu trabalhei com as pessoas que tinham problemas psicológicos. E o interessante lá é que podiam ficar bem mais tarde, juntas. Aqui nos nossos lares não, pois os idosos têm de estar todos na cama. E não têm acesso à televisão. Lá eu percebi que eles querem a televisão. Eles querem por exemplo as telenovelas. Eles adoram ver as telenovelas. Eles querem ver os programas...

[12:38 - 12:49]

- Investigadora: Então o que a senhora acha é que esse tipo de interação visual pode ser interessante para as pessoas?

- PQ3: Sim. Eu acho que sim.

[12:49 - 13:00]

- PQ9: De ver uma personagem ou alguém ali a falar pode ter um engajamento mais emotivo. Olha, eu posso usar também o exemplo da minha mãe. Eu também cuidei dela. Eu questionava tudo aos médicos nas questões de saúde que eu tinha dúvida. Eu cobrava informação deles. E a minha médica dizia assim: “ela quer saber demais”. Mas eu questionava, pois é uma coisa que nós também temos de perceber. Saber mais para poder lidar com a situação. Porque se nós não perguntamos, não estamos informados, não sabemos lidar com a situação. Não é?

[13:56 - 14:09]

- PQ12: Mas, numa situação de Alzheimer, as vezes nem os médicos dão bons conselhos.

- PQ4: Exatamente. Tens que ter tu própria o conhecimento. Aí, percebendo o que é bom para ele, o que é mau. Nosso sistema de saúde como está? Às vezes nem médico há. Porque os médicos, sabe? Não informam o doente, não dizem o que é que deve ser. Por exemplo, eles precisam de uma especialidade. Não a têm, porque nem sequer conhecem o médico das especialidades.

[14:47 - 15:00]

- Investigadora: O senhor queria comentar? Pode falar. Eu tive duas experiências. Eu fui transportador de pessoas doentes. Pessoas velhas, que eu gosto de chamar de velhinho.

[15:00 - 15:12]

PQ8: Então, foram cinco anos que eu transportei pessoas doentes. Eu tive duas experiências com algumas pessoas do Alzheimer. Uma, conhecida. Eu usava todo dia música no carro. E até gravava música. E brincávamos qual era o cantor. Eles diziam o nome. Olha, meu forte nem está em cantar. E cantavam a música. A música melhorava muito as pessoas com Alzheimer, o homem já se transformava.

[16:11 - 16:23]

- PQ9: A música ajuda a querer viver. Mas também desperta hostilidade. É que aqui, na nossa terra não se sabe cuidar dos idosos. Temos que mudar muita coisa. Não só o país, não é a terra. Eu acho que temos que mudar nós mesmos.

[16:35 - 16:47]

- PQ11: Se nós queremos que os idosos sejam tratados de outra maneira. Ou melhor. Se nós queremos ser tratados de uma outra maneira. É isso. É isso. Nós mesmos temos que fazer algo, isso há muito tempo.

- Investigadora: Só para a gente concluir. Para finalizar aqui. Para voltar para esse exercício. Vocês acham que, dentro dessa realidade de trazer interação com as pessoas e levar os conteúdos, conseguem ver que esse tipo de solução pode ser útil?

- PQ9, PQ4, PQ12, PQ7: Sim.

- PQ12: De ter uma utilidade. De repente para passar informação. Sim. Ou até como a senhora falou.

[17:56 - 18:08]

- PQ9: Às vezes nós não temos acesso ao médico. E gostaria de ter uma informação para mostrar para as pessoas. E de repente ter ali. E de repente ter acesso como uma pessoa. É. Ter aquilo muito explicado assim. Mostrando as fontes, pode ser útil.

[18:21 - 18:32]

- PQ5: Acho que como vimos muda para as pessoas. Deveria ter várias opções de soluções. O animal: alguns gostam desta solução. Outros não gostam. Seria bom ter variedades.

Appendix P

https://www.facebook.com/story.php?story_fbid=pfbid02ZfjUEa6BcknXPYZjf2B1ysk1pb8QfRzR7w5Yzc2R3q5NFg7eqt1wUj2fRYDLwDt7l&id=100064728484850&sfnsn=mo&paipv=0&eav=AfYX16W08q9ypwKObBkAOOnS6xCP92_hcNcr15BMSAJ9luduOvMFUN2KijgOsn3N9kc&_rdr

Município de Macedo de Cavaleiros
5 de dezembro de 2023 · 🌐

Evento | Workshop EDUCAGE

No âmbito do projeto EducAGE, decorreu hoje o workshop "Ferramentas Digitais de Apoio aos Cuidadores Informais: Co-criação de uma ferramenta digital de verificação de factos", na Universidade Sénior, em Macedo de Cavaleiros.

Este projeto pretende melhorar as competências dos cuidadores de idosos, rumo à implementação de estratégias de Envelhecimento Saudável com base na transição digital da sua formação e quotidiano. Visa, ainda, implementar um conjunto de materiais e ferramentas TIC que possam ser utilizadas para melhorar as competências dos cuidadores.

Durante o Workshop foi apresentada a plataforma do projeto <https://educage-project.eu/> que dá acesso a um curso online, um livro branco digital e uma App. Os participantes dividiram-se ainda em grupos para criar um avatar para o desenvolvimento de um verificador virtual, que permite confirmar a veracidade de informações sobre a saúde e combater a desinformação.

O projeto Europeu é liderado pelo Instituto Politécnico da Guarda e tem como parceiros o Laboratório Colaborativo MORE, o Centro de Cirurgia Mínima Invasión em Espanha, o Instituto Smmelweis na Hungria, a Universidade de Hadrec Králové e a Faculdade de Medicina de Hradec Králové na República Checa.

Neste workshop implementaram-se metodologias de co-criação de um verificador digital, pela investigadora Haline Maia, do Programa Doutoral em Medidas Digitais da Faculdade de Engenharia da Universidade do Porto.

Esteve presente no encerramento deste Workshop a Vereadora do Município Macedense Sónia Salomé.

👍❤️ Você e outras 61 pessoas

1 comentário · 6 partilhamentos

