recibido: 15.04.2024 / aceptado: 22.07.2024

Defending Truth and Democracy in the Age of AI: A Framework for Empowering Voters Against Persuasion and Misinformation with AI Literacy

Defender la verdad y la democracia en la era de la IA: Un marco para empoderar a los votantes contra la persuasión y la desinformación mediante la alfabetización en Inteligencia Artificial

Zeynep Arda
Izmir University of Economics

Lale Başarır
Izmir University of Economics

Referencia de este artículo

Arda, Zeynep and Başarır, Lale (2024). Defending Truth and Democracy in the Age of AI: A Framework for Empowering Voters Against Persuasion and Misinformation with AI Literacy. *adComunica. Revista Científica de Estrategias, Tendencias e Innovación en Comunicación*, n º 28. Castellón de la Plana: Departamento de Ciencias de la Comunicación de la Universitat Jaume I, 115-142. DOI: http://dx.doi.org/10.6035/adcomunica.8020.

Keywords

Artificial Intelligence; Intelligence Augmentation; Deepfakes; Critical Thinking; AI Literacy; Misinformation; Persuasive Communication.

Palabras clave

Inteligencia Artificial; Inteligencia Aumentada; *deepfakes*; razonamiento analítico; alfabetización en Inteligencia Artificial; desinformación; comunicación persuasiva.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Abstract

Witnessing the rising impact of Artificial Intelligence (AI) in a post-truth environment, our senses' credibility wanes as the distinction between the real and the imagined becomes increasingly hazy. This study examines AI-generated content and deepfakes' effects on voters by analyzing their use in Turkish and American politics over the past decade. The ease and speed of creating fake news with AI necessitate constant verification, yet regulation is often lacking, heightening daily anxiety about reality. With the EU's imminent AI law, this article highlights the dangers of media advocacy (the strategic use of media to influence public opinion) and misinformation (spreading incorrect or misleading information). Through a meticulous literature review, press release analysis, and case studies, the study assesses their possible persuasive impact on voters, elections, and democracy. It identifies the associated risks and proposes a framework to empower voters using technology to reinforce "truth filters," Intelligence Augmentation, and "AI literacy." Two concepts that were developed side by side, but with different intentions, while AI aimed to create digital intelligence, Intelligence Augmentation focused instead on using technology to enhance the human capacity.

Resumen

Al presenciar el creciente impacto de la Inteligencia Artificial (IA) en un entorno de postverdad, la credibilidad de nuestros sentidos disminuye a medida que la distinción entre lo real y lo imaginado se vuelve cada vez más difusa. Este estudio examina los efectos del contenido generado por IA y los *deepfakes* (ultrafalsos) en los/las votantes, analizando su uso en la política turca y estadounidense en la última década. La facilidad y rapidez para crear noticias falsas con IA requiere una verificación constante, pero la regulación a menudo es insuficiente, lo que aumenta la ansiedad diaria sobre la realidad. Con la inminente ley de IA de la Unión Europea, este artículo destaca los peligros de la defensa mediática (el uso estratégico de los medios para influir en la opinión pública) y la desinformación (la difusión de información incorrecta o engañosa). A través de una revisión exhaustiva de la literatura científica, del análisis de comunicados de prensa y estudios de caso, este trabajo evalúa su posible impacto persuasivo en los/las votantes, las elecciones y la democracia. A la vez, identifica los riesgos asociados y propone un marco para empoderar a los votantes utilizando tecnología para reforzar los "filtros de verdad," la Inteligencia Aumentada y la "alfabetización en IA." Dos conceptos que se desarrollaron paralelamente, pero con diferentes intenciones: Pues, mientras que la IA buscaba crear inteligencia digital, la Inteligencia Aumentada se centraba en utilizar la tecnología para mejorar la capacidad humana.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Authors

Zeynep Arda [zeynep.arda@ieu.edu.tr] is a Graphic/Interaction Designer and an Associate Professor at the Visual Communication Design Department, Izmir University of Economics, Turkey. She received her PhD degree in Communication Sciences at Universidad Jaume I, Castellón, Spain, with her dissertation "Image Becomes Identity 2.0." She holds an MA in Interaction Design (Domus Academy, Italy) and MFA in Graphic Design (Bilkent University, Turkey).

Lale Başarır [lale.basarir@izmirekonomi.edu.tr] has a PhD in Architecture from Izmir Institute of Technology, Turkey. Her research project ArchiRobie (Artificial Intelligence(AI)Architect) continues in the Izmir University of Economics where she is an Associate Professor in the Department of Architecture. She received her MSc degree from Architectural Design Computing Program at Istanbul Technical University (ITU), and her BArch in Architecture from Gazi University.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

1. Introduction

The rise of AI presents a transformative force across various sectors. However, its potential for generating deceptive content poses a significant threat to our ability to discern truth. In a "post-truth" setting, defined by the priority of emotions and opinions above objective facts, AI-generated content fuels anxieties about reality and when combined with the speed of news diffusion on social media the situation quickly turns to alarming.

The scale of the concerns in the field of communication range from being a fraud victim on a personal level, to political persuasion by deepfakes on the level of mass communication. The pursuit of persuasion through two kinds of media content; news by means of media advocacy and entertainment/education, is of particular concern when AI can be used to create fake content that influences public opinion, much faster than the verification/disclamation of such messages. The strategic use of media agencies with the deliberate intent of influencing the public opinion, especially on the voters when the elections draw near constitutes an absolute example of media advocacy (Rossini et al, 2018).

In fact, this kind of media advocacy is the theme of the dystopian future that Tegmark narrates in Life 3.0 (2017: 11). Recent research on social media carried out by an MIT team challenges conventional beliefs about the dissemination of false news, and reveals that they spread more extensively and rapidly than truth online, mostly because of their "novelty." Despite expectations that network structures and individual traits of sharers would disfavor false news, the research tells us that the opposite is observed (Vosoughi et al, 2018). It scrutinizes the levels of false and deepfake content in news and discusses the precautions in the range from critical media literacy to government regulations. Misinformation threatens democratic processes by affecting public opinion, influencing elections, and inciting violence, requiring a multifaceted response that includes fine-tuning algorithms for accuracy, reinforcing news literacy among users, and enhancing social media content moderation through advanced machine learning techniques (Aral, 2020). Following a comparative analysis of the precautions against the threats of AI, we propose a model of citizen exposure to various forms of generative AI to ensure familiarity and literacy of this technology. This path leads us to the concept of Intelligence Augmentation, a technology that develops in parallel with AI, as a human-centered, or humanity-centered solution in coping with the possible perils of persuasion. An example to augmenting means is detecting rumor veracity, a digital aid that improves rumor detection accuracy, aiding users in making informed decisions quickly even without user-specific data (Kim and Yoon, 2023).

DOI: http://dx.doi.org/10.6035/adcomunica.8020

1.1. Research Objectives

This study aims to explore the implications of AI-generated content and deepfakes on our ability to discern truth in a rapidly evolving digital landscape, focusing on the effects of AI-generated fake news on voters through diverse, impactful examples from Turkish and American political contexts. We seek to identify associated risks and propose a framework for potential solutions.

Our primary research question is: How does AI-generated misinformation influence voter perceptions and behaviors in democratic processes?

To address this, we:

- Analyze the use of AI technologies in Turkish and American politics over the past decade, assessing their impact on public opinion, elections, and the democratic process,
- Examine the dangers posed by media advocacy and misinformation, emphasizing their persuasive effects on audiences,
- Propose actionable solutions for empowering citizens and governments to combat AI-driven misinformation, including developing "truth filters" and comprehensive AI literacy programs,
- Stress the importance of strengthening defenses against these threats and raising barriers to misinformation.

Through this study, we aim to contribute to the discourse on AI, truth, and democratic integrity in the post-truth era, offering practical solutions to mitigate risks and empower individuals, and societies in navigating this complex landscape.

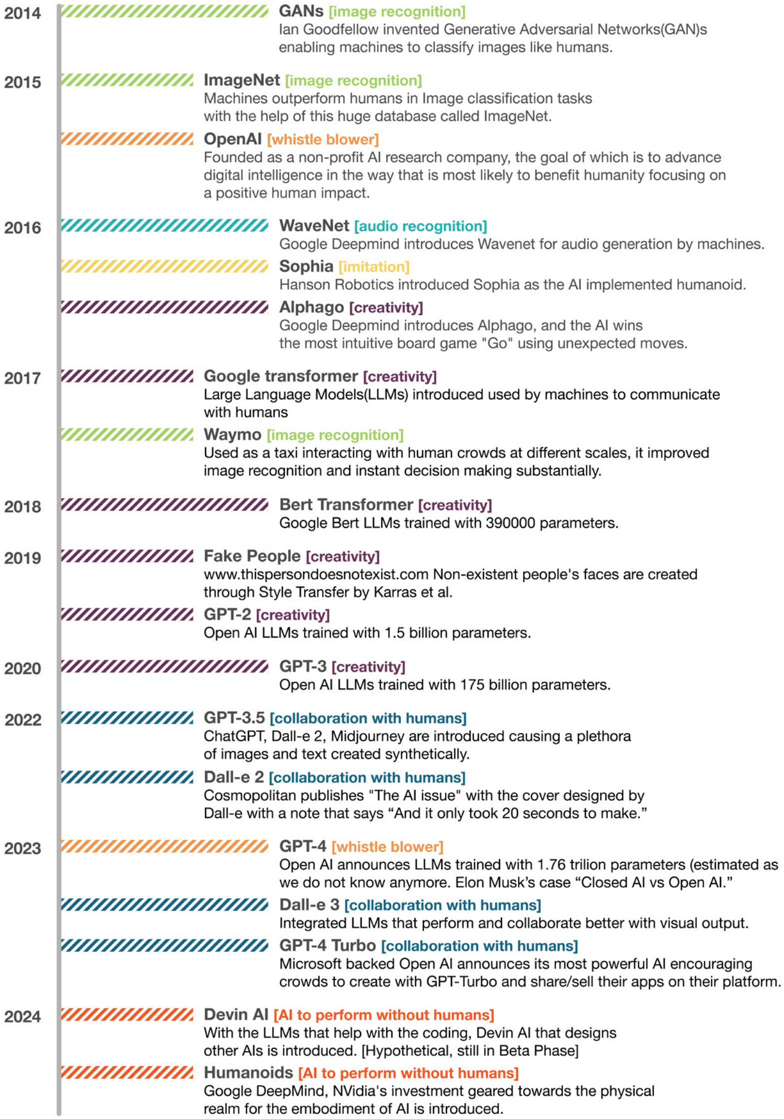
1.2. Methodology

This research displays a background to explore the complex relationship between artificial intelligence (AI) and truth in communication. The review covers sources illustrating the current landscape along with an outline of foundational concepts of AI in communication, tracing its historical context and evolution, meticulously going through the academic literature and the news media, as well as the press releases from the leading Tech companies of the field between 2014-2024. Key areas of investigation include the relationship between intelligence, AI, and the critical thinking skills essential for voters, as well as the European Union's definitions of AI and its regulatory implications.

The article further examines how AI contributes to the erosion of truth by analyzing the effects of AI-generated misinformation on persuasion tactics and identifying communication risks associated with contemporary technologies. Therefore, the methodology focuses on strategies for empowering individuals to

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Figure 1. A Brief Timeline of Technological Developments that Gave AI Its Current Capabilities



2014

GANs [image recognition]
Ian Goodfellow invented Generative Adversarial Networks(GAN)s enabling machines to classify images like humans.

2015

ImageNet [image recognition]
Machines outperform humans in Image classification tasks with the help of this huge database called ImageNet.

OpenAI [whistle blower]
Founded as a non-profit AI research company, the goal of which is to advance digital intelligence in the way that is most likely to benefit humanity focusing on a positive human impact.

2016

WaveNet [audio recognition]
Google Deepmind introduces Wavenet for audio generation by machines.

Sophia [imitation]
Hanson Robotics introduced Sophia as the AI implemented humanoid.

Alphago [creativity]
Google Deepmind introduces Alphago, and the AI wins the most intuitive board game "Go" using unexpected moves.

2017

Google transformer [creativity]
Large Language Models(LLMs) introduced used by machines to communicate with humans

Waymo [image recognition]
Used as a taxi interacting with human crowds at different scales, it improved image recognition and instant decision making substantially.

2018

Bert Transformer [creativity]
Google Bert LLMs trained with 390000 parameters.

2019

Fake People [creativity]
www.thispersondoesnotexist.com Non-existent people's faces are created through Style Transfer by Karras et al.

GPT-2 [creativity]
Open AI LLMs trained with 1.5 billion parameters.

2020

GPT-3 [creativity]
Open AI LLMs trained with 175 billion parameters.

2022

GPT-3.5 [collaboration with humans]
ChatGPT, Dall-e 2, Midjourney are introduced causing a plethora of images and text created synthetically.

Dall-e 2 [collaboration with humans]
Cosmopolitan publishes "The AI issue" with the cover designed by Dall-e with a note that says “And it only took 20 seconds to make.”

2023

GPT-4 [whistle blower]
Open AI announces LLMs trained with 1.76 trilion parameters (estimated as we do not know anymore. Elon Musk’s case “Closed AI vs Open AI.”

Dall-e 3 [collaboration with humans]
Integrated LLMs that perform and collaborate better with visual output.

GPT-4 Turbo [collaboration with humans]
Microsoft backed Open AI announces its most powerful AI encouraging crowds to create with GPT-Turbo and share/sell their apps on their platform.

2024

Devin AI [AI to perform without humans]
With the LLMs that help with the coding, Devin AI that designs other AIs is introduced. [Hypothetical, still in Beta Phase]

Humanoids [AI to perform without humans]
Google DeepMind, NVidia's investment geared towards the physical realm for the embodiment of AI is introduced.

Source: The information used to form the timeline was developed based on a continuous follow-up and evaluation of press releases between the years 2014-2024 from the leading Tech companies working on artificial intelligence and intelligence augmentation as of March 2024.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

defend against misinformation, highlighting the role of whistleblowers in promoting transparency and outlining potential regulations and guidelines aimed at mitigating the risks posed by AI-driven misinformation.

Through this structured approach, the article aims to provide a thorough examination of the challenges posed by AI in communication and propose actionable solutions for fostering a more informed and resilient society.

2. The Definition and Brief History of Artificial Intelligence in Communication

Artificial Intelligence (AI) has been a concept in science fiction for decades, but its practical applications have become increasingly prominent in the recent years. Since its formal definition by McCarthy et al (1955; 2006) as the science and engineering of making intelligent machines, AI saw substantial progress in its subfields like machine learning, natural language processing, and computer vision in the last decade (Figure 1). This progress is seen in important milestones like the invention of Generative Adversarial Networks(GANs) in 2014 (Goodfellow et al), allowing machines to exceed human performance in image classification tasks by the year 2015.

The field saw further advancements with the introduction of WaveNet for audio generation (Oord et al, 2016), the Transformer architecture for language models (Vaswani et al, 2017) and the continued development of large language models like GPT-3 (OpenAI, 2020). These advancements extend beyond language and image generation with AI playing a crucial role in self-driving cars and achieving a superhuman performance in complex strategy games. The recent unveiling of GPT-4 by OpenAI in 2023 highlights the ongoing debate surrounding the level of openness in AI development, while Dall-e 3 showcased the ability to generate incredibly realistic images within seconds. The introduction of Devin AI (a closed model at the time of the writing of this article) in 2024 demonstrates the rapid progress in the field, highlighting AI's ability to design other AI models that are new and that were not conceived by human beings.

2.1. Intelligence, AI and the Significance of Critical Thinking for Voters

Researchers working in the field of AI made adamant efforts to define "intelligence" (Gottfredson, 1997; Buchanan, 2006; Legg, S.& Hutter, 2007). Their definitions imply that human intelligence consists of numerous skills and abilities that enable human adaptation to various levels and instances of their environment. In this research, the term "intelligence" refers to the mechanism that enables humans to connect with and adapt to their environments, shaped by a network of bounding realities. As human nature has an inherent tendency to challenge its limits, the AI research was marked by the quest for a "superintelligence" that was above hu-

DOI: http://dx.doi.org/10.6035/adcomunica.8020

man intelligence. Nonetheless, critical thinking and the filtering of AI-generated content seems to be crucial for the future of human societies and this is not yet achieved by the AI models that are in use at the moment.

As AI continues to evolve, its impact on communication and interaction with the world around us promises to be profound. And critical thinking becomes the most important component for human-beings that become the subjects and targets of AI-generated content: If the voters are impacted by a combination of media advocacy, misinformation and fake news the overall threat on democracy would end up being irreversible. Lead by these concerns this article searches for the current precautions that would encourage, strengthen and empower critical thinking in voters: What could be possible government actions to regulate the use of AI in communication, what are the roles assumed by the companies developing these technologies and whether the technology itself can be employed to help the voters combat the possible negative effects of AI-generated content.

2.2. European Union's Definition of AI for Regulating the Relevant Practices

The EU AI Act defines AI systems as "software that is developed using one or more of the techniques and approaches listed in Annex I" (Artificial Intelligence Act, 2021-23). This annex details various machine learning techniques, including deep learning, which are commonly used in AI development. This definition is aligned with the criteria for AI was formerly defined by the OECD guidelines as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."

3. Artificial Intelligence and the Erosion of Truth

Dreaming of his Difference Engine, Charles Babbage who originated the concept of a digital programmable computer, was imagining "the substitution of human labor with a machine" (Purbrick, 1993: 20). While the eventual outcome of this line of thought and the development of "machines that can think" brought us to the discussions of artificial intelligence that we have today, it would be safe to say that Babbage would not have imagined that this human labor also included the human act of lying. Umberto Eco, while he was defining semiotics, said that it is in principle the discipline of studying everything which can be used in order to lie: "If something cannot be used to tell a lie, conversely it cannot be used to tell the truth: it cannot in fact be used 'to tell at all'" (Eco, 1976: 7).

Interestingly, Danesi visited the etymological root of the word "mafia" in his discussion of Eco's theory: "Intentional lying has always had specific (marked) functions. It has, for instance, always been a part of criminal organizations, such

DOI: http://dx.doi.org/10.6035/adcomunica.8020

as the Mafia, which use falsification and confabulation to justify their legitimacy. Even the word mafia is a made-up form, appearing in an 1868 dictionary, where it is defined as "the actions, deeds, and words of someone who tries to act like a wise guy" (Danesi, 2017: 23).

Today, AI excels at manipulating data, creating realistic simulations, and generating persuasive narratives, almost as the mafia would do some decades ago to blackmail its victims. Though AI is not the sole responsible, this capability allows for the creation of a wide spectrum of misinformation artifacts from "astroturfing" to "unverified content" and "deepfakes," highly realistic digital forgeries that can be used to disinform and erode public trust. The ease and speed of generating fake news or deepfakes further amplify the issue, forcing us to constantly question the authenticity of information encountered online.

In his book Life 3.0, Max Tegmark tells the cautionary tale of Prometheus, the super AI created secretly inside a tech company by the "Omegas" (2017: 14). Though this team of engineers keep Prometheus contained in its designated space by not allowing it to access internet, once they decide that the AI is ready to program other AI systems, they try it on the "human intelligence tasks" on Amazon Mechanical Turk. To prepare for the tasks lying ahead, Prometheus learns everything on the local copies of Wikipedia, the Library of Congress, Twitter, a selection from YouTube, much of Facebook. On the first day of testing its "intelligence" on these tasks, Prometheus is so fast and successful that the Omega team earns millions to finance the second step of their project. Next, they build a media company with Prometheus, starting with a network platform with animated movies and TV series, completely developed by AI and at almost zero cost, and once they cover this step as well, they move on to "the next step of their audacious plan: taking over the world. Within a year of the first launch, they had added remarkably good news channels to their lineup all over the globe. As opposed to their other channels, these were deliberately designed to lose money, and were pitched as a public service. In fact, their news channels generated no income whatsoever: They carried no ads and were viewable free of charge by anyone with an internet connection" (Tegmark, 2017: 26).

The second phase of Prometheus' news strategy was persuasion: "Their plethora of channels catering to different groups still reflected animosity between the United States and Russia, India and Pakistan, different religions, political factions and so on, but the criticism was slightly toned down, usually focusing on concrete issues involving money and power rather than on ad hominem attacks, scaremongering and poorly substantiated rumors" (Tegmark, 2017: 27).

DOI: http://dx.doi.org/10.6035/adcomunica.8020

3.1. Misinformation and Persuasion in the Era of AI

Even without the interference of AI-developed media, the current state of news media seems to be doing more than what traditional journalism offered: Persuasion. Traditional newscasters were "not supposed to show emotion when delivering the news," (Schudson 2001: 150; Richards and Rees, 2011) however, starting with Fox News, they started providing a type of entertainment disguised as news, as their anchors were paid to convey negative emotions that underpinned their delivery of the content (Spitale, 2022: 122). Some scholars consider this "affective turn" as necessary and acceptable to engage the audience in the news (Barnidge, 2018; Parks, 2021; Pajnik, 2024). However, feeding opinions to voters instead of giving them unbiased information so that they can form their own opinions is called persuasion. The problem with persuasion, be it through AI-fabricated fake news, or through the attitudes of new age journalists, is that it prevents the audience from making informed decisions: "Invoking negative emotions in the audience triggers their fight-or-flight response, which overrides critical thinking, both on-screen and off" (Spitale, 2022: 123). This is the classical utilization of the Aristotelian pathos, using an emotion-based argument to play upon the fears and desires of an audience.

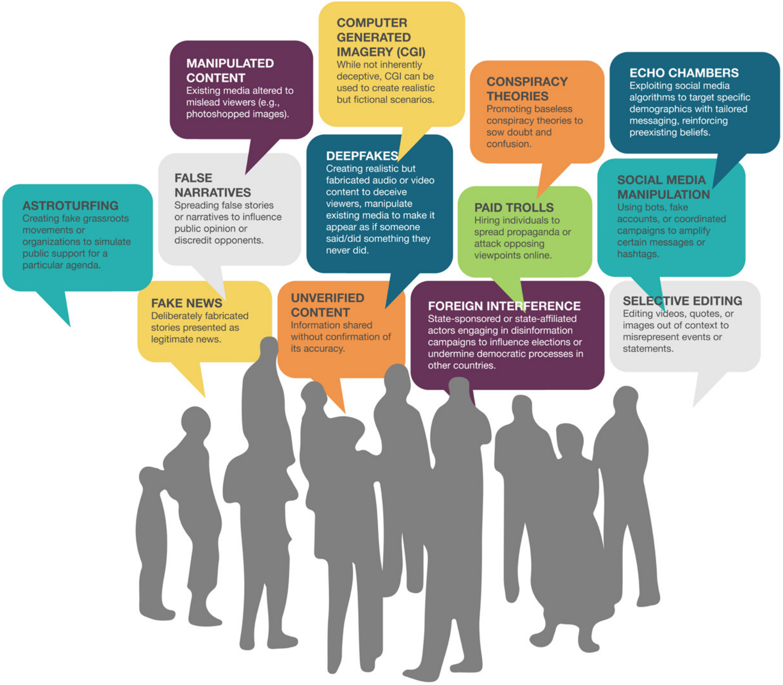
Such possibilities for persuasion require a reevaluation of the definition of audience and a sort of overlapping of modernist and postmodernist theories of media: On one hand, the speed of production and distribution of AI-generated imagery create an audience that can be considered as "gullible" in a sense similar to that of modernist media theories. On the other hand, the way all tech-savvy human-beings being able to sort to AI to create such imagery coincides with the post-modernist theories that place the audiences as "prosumers" (Toffler, 1981: 11). While digital information, verbal or visual, can be produced by anyone and there are almost no mechanisms of verification before it floods the social media, its potential manipulation for political persuasion or disinformation is heightened. Still human factors appear to excel in these manipulations, as a meta-analysis on persuasion outcomes found the AI agents to be as persuasive as humans in the overall persuasion outcomes, while AI was found to be less effective at shaping behavioral intentions (Huang and Wang, 2023).

At this point, it is important to differentiate between various types of misleading content and understand the spectrum of misinformation. Figure 2 includes brief descriptions of the confabulation techniques available in today's media.

These tactics are employed by various actors for persuasion, including political parties, governments, foreign entities, and ideological extremists, to manipulate public opinion, sway elections, and advance their agendas. AI is employed in the creation or manipulation of some of these confabulation methods, like deepfakes or echo chambers, however it should be kept in mind that the current technology requires human input for all such uses of Artificial Intelligence.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Figure 2. The Spectrum of Misinformation


COMPUTER GENERATED IMAGERY (CGI) While not inherently deceptive, CGI can be used to create realistic but fictional scenarios.
MANIPULATED CONTENT Existing media altered to mislead viewers (e.g., photoshopped images).
CONSPIRACY THEORIES Promoting baseless conspiracy theories to sow doubt and confusion.
ECHO CHAMBERS Exploiting social media algorithms to target specific demographics with tailored messaging, reinforcing preexisting beliefs.
DEEPFAKES Creating realistic but fabricated audio or video content to deceive viewers, manipulate existing media to make it appear as if someone said/did something they never did.
FALSE NARRATIVES Spreading false stories or narratives to influence public opinion or discredit opponents.
ASTROTURFING Creating fake grassroots movements or organizations to simulate public support for a particular agenda.
PAID TROLLS Hiring individuals to spread propaganda or attack opposing viewpoints online.
SOCIAL MEDIA MANIPULATION Using bots, fake accounts, or coordinated campaigns to amplify certain messages or hashtags.
FAKE NEWS Deliberately fabricated stories presented as legitimate news.
UNVERIFIED CONTENT Information shared without confirmation of its accuracy.
FOREIGN INTERFERENCE State-sponsored or state-affiliated actors engaging in disinformation campaigns to influence elections or undermine democratic processes in other countries.
SELECTIVE EDITING Editing videos, quotes, or images out of context to misrepresent events or statements.

Source: compiled by authors.

3.2. Communication Risks with Current Technologies and AI

While social media platforms and existing technologies like Photoshop, have facilitated the spread of misinformation even before the widespread use of AI in the past couple of years, AI is commonly presented as a unique threat by media outlets heightening the fears of the audiences (Cole, 2017; Parkin, 2019; Ellery, 2023; McCarthy, 2023; Ulmer and Tong, 2023; Harford, 2024; Spring, 2024). AI's ability to automate content creation and generate highly realistic deepfakes makes it significantly easier to produce and disseminate deceptive material even though fake news, unverified, or manipulated content was already used persuasively in the past years.

In 2021, the release of DALL-E, a transformer-based pixel generative model, followed by Midjourney and Stable Diffusion, marked the emergence of practical high-quality Artificial Intelligence art from natural language prompts. The threat

DOI: http://dx.doi.org/10.6035/adcomunica.8020

was already reported by mainstream media in 2017, as the journalist Samantha Cole introduced the risk of deepfakes with the following sentence: "There's a video of Gal Gadot having sex with her stepbrother on the internet." The video was indeed a deepfake that was placing Gadot's face on a porn performer's body, and it had only served to pour oil on the flame: "We are on the verge of living in a world where it's trivially easy to fabricate believable videos of people doing and saying things they never did" (Cole, 2017). The ease of creating them expanded the hinterland of deepfakes in the years that followed, making them a widespread concern for all citizens: Young school girls in the small Spanish town Almendralejo, whose deepfake nude images were created by the ClothOff AI were suffering panic attacks and refusing to go to school. On the other side of the ocean, at the Westfield High School in New Jersey, other young girls were targeted by similar images created by the same app. Moreover, the anonymity of the people behind the app were carefully guarded by an entirely fake, AI-generated person who claimed to be their CEO (Safi, Atack and Kelly, 2024).

This cheapening of falsification made it seem as if deepfakes were posing the bigger threat on the truth, nonetheless, two analysts explaining Russia's disinformation and persuasion strategies thought otherwise: Our conventional understanding of propaganda was that its messages should be mostly true, and even if they are not, they should be believable and consistent. But Russian media channels, websites and social media accounts for rent would post anything, without even considering whether they are true, false or believable. Their strategy called the "firehose of falsehood" only considered "speed, relevance and volume" as the significant parameters for social media-fueled propaganda (Paul and Matthews, 2016). The impact of this strategy can go beyond an evaluation of truth or skepticism, and lead to a complete disengagement from such effort of looking for the truth instead. The rapid circulation of similar viewpoints from various sources can make a convincing impact, and even if the information is false, it can still saturate social and conventional media with distractions and negativity, deterring news consumers (Harford, 2024). And of course, all this false information can easily be generated by AI.

Investigating a brief history/timeline of political persuasion in the last decade ranging from election fraud (personalized messages on FB) highly-disseminated fake news, deepfakes and unverified/manipulated content, in light of how they work on audiences in their persuasive capacities, could help in identifying the specific criteria for defending the truth.

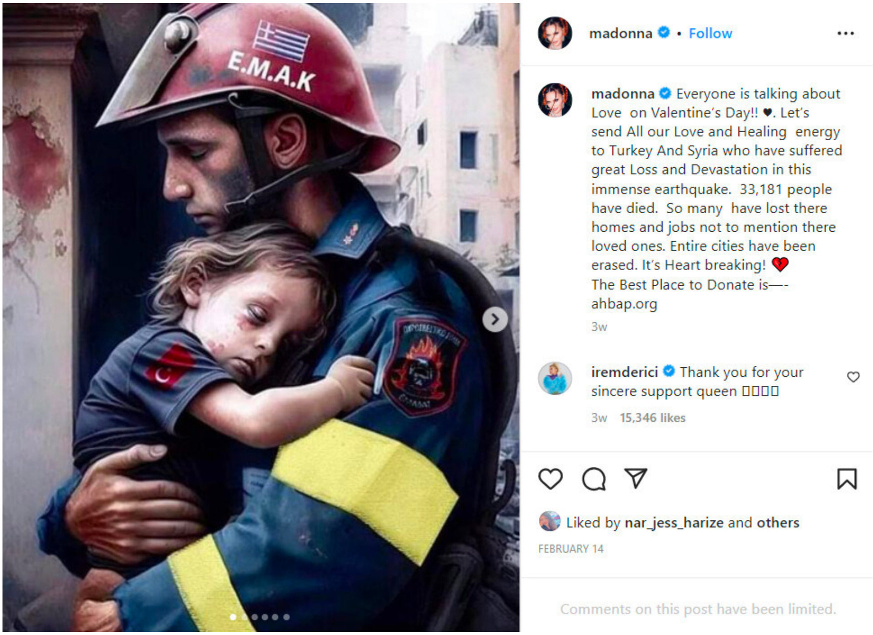
The Cambridge Analytica scandal, where Facebook data of 87 million users was harvested to target voters with personalized political messaging, served as a stark reminder of the potential for social media to be weaponized for misinformation campaigns. While the harvesting and misuse of the data took place before the 2016 U.S. Presidential Elections, the scandal made the news in March 2018, ending up

DOI: http://dx.doi.org/10.6035/adcomunica.8020

with the Facebook founder Mark Zuckerberg testifying before the Senate and House committees (Confessore, 2018). This scandal served as an early warning in the misuse of public trust for political persuasion and foreign interference.

In February 2023, a couple of days after a devastating earthquake shook Turkey and Syria, an image of a Greek social worker comforting a Turkish child started circulating on social media (Harize, 2023). Though the image meant well and displayed solidarity between the Greek and the Turkish people, which was already taking place in the ten Turkish provinces severely affected by the earthquake, it lacked authenticity. The probability of a rescued child wearing a t-shirt conveniently carrying a Turkish flag on its arm was already suspicious, but it did not keep the AI-generated image from being shared notoriously by wide audiences, moved by the warmth it displayed – including Madonna who shared the image on her Instagram account (Figure 3). As a visual that connotes the friendship of the two neighboring countries, it was an example of a moment that was never real, but it did highlight the emotional impact of visuals in general and AI-generated ones in particular.

Figure 3. AI-Generated Image of a Turkish Child and Greek Rescuer on Madonna's Instagram



madonna Everyone is talking about Love on Valentine's Day!! ♥. Let's send All our Love and Healing energy to Turkey And Syria who have suffered great Loss and Devastation in this immense earthquake. 33,181 people have died. So many have lost there homes and jobs not to mention there loved ones. Entire cities have been erased. It's Heart breaking! ❤️ The Best Place to Donate is—- ahbap.org

3w

iremderici Thank you for your sincere support queen □□□□

3w 15,346 likes

Liked by nar_jess_harize and others

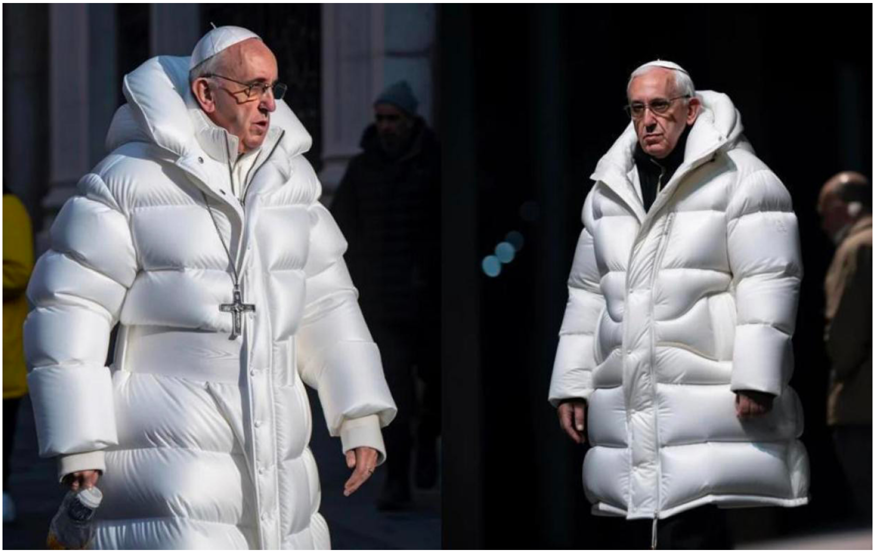
FEBRUARY 14

Comments on this post have been limited.

Source: compiled by authors.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Figure 4. Pope Francis in a Puffer Jacket



Source: compiled by authors.

In March 2023, this time it was an image of Pope Francis in a stylish puffer jacket and a jeweled cross necklace that did the rounds (Figure 4). A satirical news story with the AI-generated image of the Pope went viral, once again, highlighting the ease of manipulating images (Ellery, 2023). Though these two images mentioned did not create any apparent political percussions, deep down they were subliminally and sneakily etching ideas in the minds of the viewers. These instances were soon followed by a political campaign by Ron DeSantis, where the Florida governor's media team interspersed three authentic photographs showing Donald Trump and Dr. Fauci together at various press conferences, with three AI-generated images showing them hugging and embracing. These images were fact-checked and scrutinized by AFP resulting in their identification as "false" (McCarthy, 2023). This incident underscored the vulnerability of political discourse to AI-generated manipulation, as they demonstrate "how the 2024 Republican White House contenders have elevated their war of words into the AI-driven social media arena, interspersing fact with fiction" (Ulmer and Tong, 2023).

A more recent manipulated content came from the other side of the world, and as the Israel-Hamas war toughened, a video allegedly showing an explosion received hundreds of thousands of views on social media posts that falsely claimed it as showing the conflict in the Middle East. In fact, the video showed a 2016 explosion

DOI: http://dx.doi.org/10.6035/adcomunica.8020

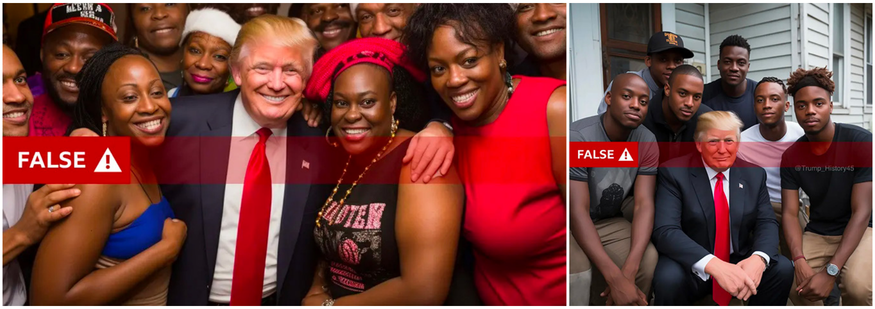
at a fireworks market in Tultepec, Mexico. It was first spotted on Tiktok in October, 2023, with a headline that reads "Palestine vs Israel is getting heated, is this a sign that World War III is happening?" in Indonesian on the top. The video was falsely contextualized as it showed an irrelevant explosion that happened years before the Israel-Hamas war (AFP Indonesia, 2023).

In March 2024, as Turkey's local elections approached, polls indicated a tight race between Erdogan's AK Party and the current mayor of Istanbul. An AI-generated video of Mayor Ekrem Imamoglu praising Erdogan circulated on social media, sparking concerns over fake news. Mainstream media, largely government-controlled, failed to verify the videos' authenticity. The fact that Erdogan had previously used a fake video against his opponent Kilicdaroglu in the general elections in 2023, raised alarms (Jones, 2024).

Around the same time in the United States, supporters of Donald Trump were busy circulating AI-generated fake images of black voters to sway African Americans to vote Republican, as discovered by BBC Panorama (Figure 5). These deepfakes depicted black individuals endorsing the former president, despite no direct evidence linking them to Trump's campaign. This manipulation served a strategic narrative that portrayed Trump as popular within the black community, and one of the voters, who created these images using AI admitted their inaccuracy. Despite their misleading nature, some viewers mistook these images for real, accentuating the potential impact of such disinformation on public perception (Spring, 2024).

Currently, the overall impact of fake news and AI-generated images is mostly created by human beings, however, as the versions of AI systems that can perform without human intervention are being beta tested, talking about critical media literacy has to involve the work of both of these factors. The way fake news affects the audience's perspective is in deactivating critical thinking, which takes place slower than black/white thinking that these images provoke. Black/white thinking

Figure 5. AI-Generated Image of Trump with Black Voters



Source: compiled by authors.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

works on the polarities and creates an immediate response, while critical reading of such images requires their contextualization.

False information has exacerbated the erosion of trust in media, politics, and established institutions globally. While emerging technologies such as artificial intelligence (AI) could make things even worse, AI and AI literacy can also offer potential solutions to counter misinformation (Cassauwers, 2019). Some of these solutions are already available and they provide the metrics for assessing the informativeness of articles, testing their source quality and credibility, past performance, author expertise, diversity and tone (Van Der Lans, 2021).

4. Defending the Truth and the Democracy: Empowering Voters for Self Defense with Artificial Intelligence and Intelligence Augmentation Literacies

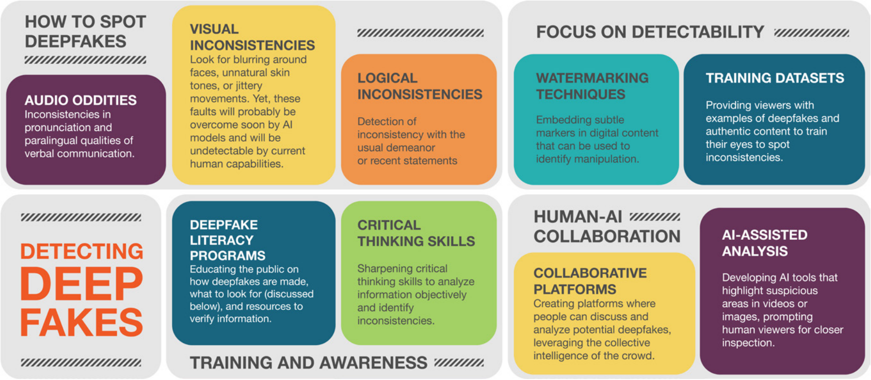
4.1. AI Literacy and Familiarization of Users with AI

Though a long tradition of social science and psychology suggests that human beings are gullible, that they are easily persuaded via messages, if they are "reasonably well adapted," they would be vigilant toward such communication (Mercier, 2017). In this section, we will discuss what the criteria for AI adaptation could be, so that they would be aware of the perils yet would choose to use this technology in an ethical manner. In this study, we have shown that the media messages that monger the dangers associated with AI in a way that exempts the human factors involved in such harmful uses, but censorship or labeling of AI-created media may hinder the fact that humans can excel at detecting fake content. Researchers suggest a reevaluation of precautionary measures toward content generation technologies (Groh et al, 2021). Their findings suggest that people should be exposed to manipulated content, in order to be better trained in detecting fake information, and that this ability would only improve with exercise.

While these encounters would make human beings "AI literate," the AI would also benefit from such interactions. Alphago's success at playing Go in 2016 was a milestone in demonstrating that the AI can indeed make connections that it is not programmed to make, hence through this interaction, it can deduce new use scenarios that we did not foresee and propose novel and creative uses for fake detection. Meanwhile, citizens attempt to cope with fake news and AI generated content, such as the Trusted Web Foundation (TWF), co-funded by the European Union in the Netherlands, support the involvement of AI in fake detection. Considering the necessity of trust, transparency, and accountability across all online domains, including news, social media, and e-commerce, TWF formulated the "timestamping" of content for enabling the verification of information sources, ownership confirmation, and conflict resolution (TWF, 2021).

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Figure 6. Detecting Deep Fakes and Augmenting Human Intelligence



HOW TO SPOT DEEPFAKES

AUDIO ODDITIES
Inconsistencies in pronunciation and paralingual qualities of verbal communication.

VISUAL INCONSISTENCIES
Look for blurring around faces, unnatural skin tones, or jittery movements. Yet, these faults will probably be overcome soon by AI models and will be undetectable by current human capabilities.

LOGICAL INCONSISTENCIES
Detection of inconsistency with the usual demeanor or recent statements

FOCUS ON DETECTABILITY

WATERMARKING TECHNIQUES
Embedding subtle markers in digital content that can be used to identify manipulation.

TRAINING DATASETS
Providing viewers with examples of deepfakes and authentic content to train their eyes to spot inconsistencies.

DETECTING DEEP FAKES

DEEPFAKE LITERACY PROGRAMS
Educating the public on how deepfakes are made, what to look for (discussed below), and resources to verify information.

CRITICAL THINKING SKILLS
Sharpening critical thinking skills to analyze information objectively and identify inconsistencies.

TRAINING AND AWARENESS

HUMAN-AI COLLABORATION

COLLABORATIVE PLATFORMS
Creating platforms where people can discuss and analyze potential deepfakes, leveraging the collective intelligence of the crowd.

AI-ASSISTED ANALYSIS
Developing AI tools that highlight suspicious areas in videos or images, prompting human viewers for closer inspection.

Source: compiled by authors.

4.2. Deepfakes and the Right to Information

The human ability to recognise fakeness is deeply rooted in the recognition of faces and facial expressions, and the emotional evaluations rather than rational ones. A study investigating the cognitive and emotional engagement of 23 healthy individuals' in a visual discrimination task that employed real and deepfake human faces expressing positive, negative, and neutral emotions found that the participants discriminated between the real and synthetic faces for the former. Despite its limited sample, this study revealed statistically significant activations in specific brain areas depending on the authenticity and emotional content of the stimuli (Tarchi et al, 2023).

As deepfakes are constantly evolving, the human capacity to detect deepfakes can be improved by constant exercise of critical thinking and social awareness coupled with AI's analytical capabilities. Confirming that the emotional and cognitive processes worked together in perceiving fakeness (Tarchi et al, 2023), we have compiled the information in Figure 6 to propose some of the avenues where human abilities can be enhanced:

Both UNESCO and the European Union regard media education as part of fundamental and universal rights as important as freedom of expression and the right to information, playing a key role in the construction and preservation of democracy. Media and information literacy (MIL) encompasses a range of skills essential for navigating the modern information landscape effectively and responsibly. MIL involves competencies such as information searching, critical evaluation, ethical use of content, and combating online issues like hate speech and cyberbullying. It

DOI: http://dx.doi.org/10.6035/adcomunica.8020

also includes understanding one's online rights and promoting values like equality and free expression through media engagement. These capacities are crucial for individuals of all ages and backgrounds to maximize benefits and minimize risks in the digital age (UNESCO, 2022). However, in this report the AI is only addressed twice and only as part of the digital communication landscape and not as a potential threat to the communication rights.

4.3. Utilization of AI-Powered Tools for Detection

One of the early definitions identifies media literacy as "the ability to create personal meaning from the visual and verbal symbols we take in every day from television, advertising, film, and digital media," and emphasizes the need for the audiences to be "critical thinkers who can understand and produce in the media culture swirling around them" (Adams and Hamm, 2001). Current research focusing on deepfake detection places human beings' critical abilities above that of the machine learning models, identifying the human groups, as achieving high accuracy especially in deepfakes that involve manipulations of facial features. However, human beings hesitate and second-guess while the machines decide once and for all, therefore the most accurate results are obtained when the machine and human judgments are used together (Groh et al, 2021, Hassani et al, 2020). These findings suggest that the detection of truth should be multi-layered and that it should include the AI and human beings as collaborators instead of opponents.

The regulation paves the way for the development of AI-powered tools capable of detecting and flagging deepfakes. These challenges include potential biases within AI systems used for detection and the potential for an "arms race" between deep fake creators and detectors. The human-AI collaboration may come in various forms, all of which require more interest and involvement of human beings in AI. While many companies employ AI in media forensics and truth-verification software to combat fake news in media, the concerns of the developers are twofold: Some defend the significance of evidence-based conclusions while others claim that public skepticism would be a better defense than authentication tools (Parkin, 2019).

4.4 Human-AI Collaborations: Intelligence Augmentation

This collaborative sense of Human-AI interaction is supported by another school of thought which suggests that instead of being a replacement for the human mind, AI can be used for "intelligence augmentation" (Hassani et al, 2020). Shorthanded as IA, this standpoint and such artificial awareness could be useful in handling the potential risks of technology, including those related to the media and informative rights of citizens.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

The term intelligence augmentation was coined by Douglas Engelbart, who first discovered the importance of computer technologies in bootstrapping human creativity and capabilities (Engelbart, 1962). His concept of IA, “halfway between the entirely human and entirely automated capabilities” that aimed at improving the efficiency of human intelligence evolved side by side with AI. Despite the popularity of AI, IA competed silently and not surprisingly, non-technologically literate individuals are unlikely to have heard about it (Hassani et al, 2020).

The shift of focus from AI to IA allows us to understand the crucial progress that can be achieved by integrating the two forms of intelligence, the organic and the synthetic, rather than pushing them to compete. The two current examples of IA, Grammarly and Photoshop illustrate two instances where AI can be used to help the individual excel by providing the necessary tools for the efficiency of the human agent and decreasing the time cost (Hassani et al, 2020). Both these technologies can play a crucial role in combating the fake factors in informative media, as they would allow the human beings to make informed decisions with AI providing the data set and the IA helping the humans make better predictions using that data.

Such paradigm shifts can also allow us to see the impact of AI/IA in a different light. The human brain training and exercising to detect AI or human generated misinformation and if all the precautions and mechanisms that are suggested in this section would be put to use in this sense, the audiences will become better critical thinkers to arm themselves against ill-willed fake acts and the skeptical/informed attitude that they will develop as a result would end up contributing positively to democracy on the global level.

4.5. The Impact of the Whistleblower

Bill Joy, a computer scientist, expresses deep concern about the potential dangers of rapidly advancing technologies like robotics, genetic engineering, and nanotechnology (2001). He worries that these advancements could make humans endangered or even obsolete. The text revolves around a conversation with Ray Kurzweil, a renowned inventor, who predicts a future where humans merge with machines to achieve near immortality. Joy is concerned by the little attention paid to possible negative consequences, even though he recognizes the utopian potential in this vision. Joy presents a dystopian scenario from the Unabomber Manifesto (Kaczynski, 1995), which depicts a future where intelligent machines control everything. Humans become either dependent or irrelevant, with an elite few holding all the power. This scenario fuels Joy’s anxieties about the future and the potential misuse of these powerful technologies.

Kurzweil advocates for a measured approach to the development of advanced technologies (2001). He acknowledges possible risks and worst-case scenarios theoretically. Still, he contends that the most dependable, if not perfect, route to maximizing the potential advantages of these technologies while minimizing their potential drawbacks is to pursue relatively unrestricted scientific exploration (Kurzweil, 2001).

Whistleblowers like Frances Haugen, who exposed Facebook's internal struggles with misinformation, play a crucial role in raising awareness about these issues (Wong, 2021).

Elon Musk has left OpenAI announcing that he was working on Tesla's AI and implied a conflict of interest if he continued to be part of the organization. However, his latter comments indicated that he lost his faith in the once not-for-profit platform OpenAI and that he feared several aspects of AI gaining powers.

Geoffrey Hinton, who is the brain behind ImageNet, that is one of the most impactful leaps of Artificial Intelligence (Krizhevsky et al, 2017), reportedly left Google in 2023 to be able to more freely discuss its potential dangers. He expressed a desire to speak out without being constrained by potential impacts on Google itself, which he acknowledged has acted responsibly in AI development. Hinton is generally considered a leading figure who helped spark the recent advancements in AI, particularly with deep learning techniques. While he acknowledges potential risks with AI, his focus tends to be on ensuring its development is done responsibly and safely.

The EU had also taken on the role of a whistleblower in 2018 when it brought about the precautions on data privacy.

4.6. Regulations and Guidelines

In response to many concerns around the developments in the AI domain, the European Union is introducing the first comprehensive AI regulation. The 2018 implementation of the EU's General Data Protection Regulation (GDPR) aims to safeguard individual privacy by controlling the collection, use, and storage of personal data. Although not specifically addressed, AI development is subject to the GDPR's rules regarding transparency, data minimization, and restrictions on automated decision-making. On the other hand, the EU AI Act, which is currently undergoing negotiations, intends to regulate the creation and application of AI systems within the EU. To do this, AI will be categorized according to risk, and applications deemed high-risk will be subject to stronger regulations. Thus, the GDPR offers a broad framework for data protection, and the AI Act seeks to expand upon it by enacting laws specifically relevant to AI.

The EU legislation aims to establish ethical guidelines for AI development and deployment. A key aspect of the regulation focuses on mitigating the risks associated with AI-generated misinformation. The paper will analyze the specific

DOI: http://dx.doi.org/10.6035/adcomunica.8020

measures proposed by the EU to address concerns about deep fakes and promote trustworthy information ecosystems.

A range of strategies for AI governance are shown in Figure 7, which also highlights several important strategies for controlling the advancement and application of AI. The Asilomar AI Principles (Tegmark, 2017: 329) and Microsoft's AI Principles offer ethical frameworks emphasizing fairness, transparency, accountability, and overall societal well-being. These are non-binding guidelines intended to steer responsible AI development. In contrast, the EU AI Act represents a legally enforceable approach, focusing on mitigating the immediate risks posed by specific AI applications. It classifies AI systems based on risk and mandates specific requirements for high-risk categories.

While both principles and regulations aim to achieve safe and beneficial AI, Hinton's research takes a different perspective. He focuses on the technical aspects of AI safety, particularly the alignment problem. This concept addresses how to ensure AI goals remain aligned with human values even if AI surpasses human intelligence. Hinton also emphasizes designing long-term reward functions that incentivize desirable AI behavior over the long term and developing interpretable AI systems to understand how they arrive at decisions.

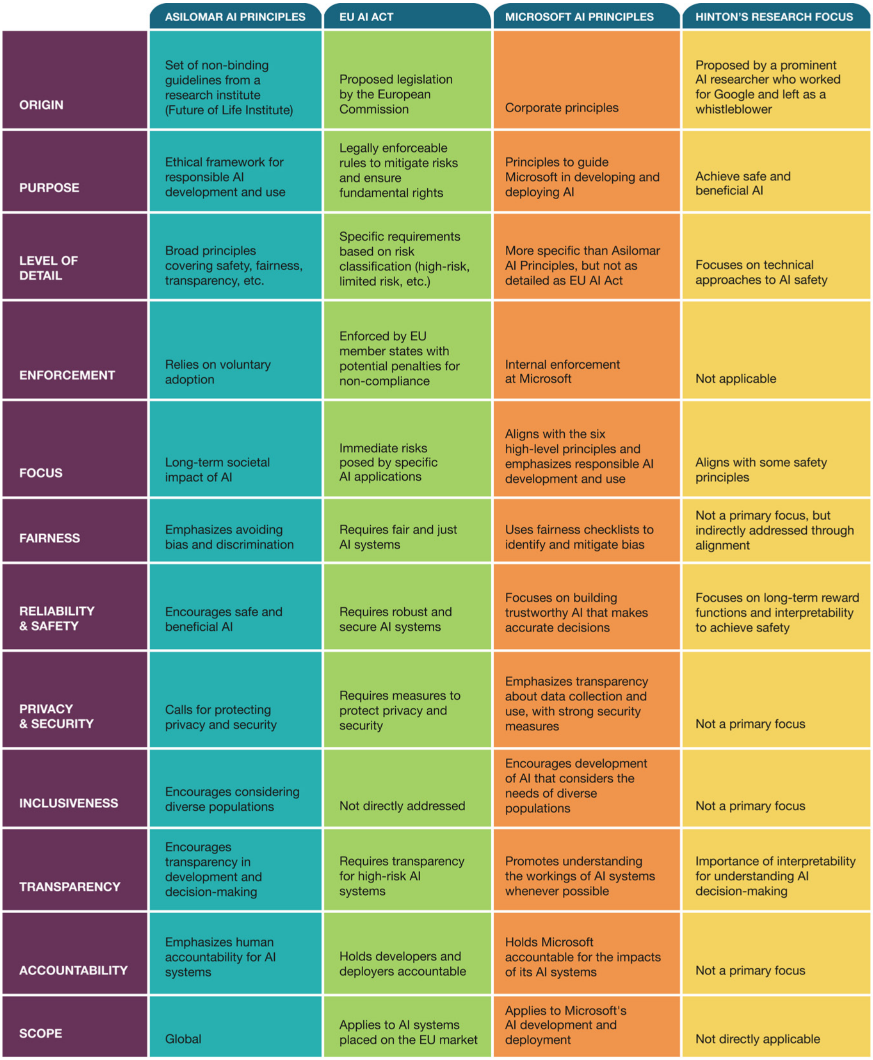
Principles and regulations provide frameworks for ethical development and risk mitigation, while Hinton's research delves into the technical challenges of achieving safe and aligned AI.

4.7. Further Research

This study was carried out to serve as a springboard for further research areas. These include the psychological impact of deep fakes on individuals and societies, the development of educational programs aimed at enhancing AI literacy, awareness of possible contributions of IA and the exploration of ethical frameworks for AI development and deployment. In order to set the pathways for further research, the basis developed by this juxtaposition of different approaches can be employed and expanded to secure the global human right to information and democracy.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Figure 7. A Comparison of AI Principles and Regulations Suggested by Entities and Institutions



	ASILOMAR AI PRINCIPLES	EU AI ACT	MICROSOFT AI PRINCIPLES	HINTON'S RESEARCH FOCUS
ORIGIN	Set of non-binding guidelines from a research institute (Future of Life Institute)	Proposed legislation by the European Commission	Corporate principles	Proposed by a prominent AI researcher who worked for Google and left as a whistleblower
PURPOSE	Ethical framework for responsible AI development and use	Legally enforceable rules to mitigate risks and ensure fundamental rights	Principles to guide Microsoft in developing and deploying AI	Achieve safe and beneficial AI
LEVEL OF DETAIL	Broad principles covering safety, fairness, transparency, etc.	Specific requirements based on risk classification (high-risk, limited risk, etc.)	More specific than Asilomar AI Principles, but not as detailed as EU AI Act	Focuses on technical approaches to AI safety
ENFORCEMENT	Relies on voluntary adoption	Enforced by EU member states with potential penalties for non-compliance	Internal enforcement at Microsoft	Not applicable
FOCUS	Long-term societal impact of AI	Immediate risks posed by specific AI applications	Aligns with the six high-level principles and emphasizes responsible AI development and use	Aligns with some safety principles
FAIRNESS	Emphasizes avoiding bias and discrimination	Requires fair and just AI systems	Uses fairness checklists to identify and mitigate bias	Not a primary focus, but indirectly addressed through alignment
RELIABILITY & SAFETY	Encourages safe and beneficial AI	Requires robust and secure AI systems	Focuses on building trustworthy AI that makes accurate decisions	Focuses on long-term reward functions and interpretability to achieve safety
PRIVACY & SECURITY	Calls for protecting privacy and security	Requires measures to protect privacy and security	Emphasizes transparency about data collection and use, with strong security measures	Not a primary focus
INCLUSIVENESS	Encourages considering diverse populations	Not directly addressed	Encourages development of AI that considers the needs of diverse populations	Not a primary focus
TRANSPARENCY	Encourages transparency in development and decision-making	Requires transparency for high-risk AI systems	Promotes understanding the workings of AI systems whenever possible	Importance of interpretability for understanding AI decision-making
ACCOUNTABILITY	Emphasizes human accountability for AI systems	Holds developers and deployers accountable	Holds Microsoft accountable for the impacts of its AI systems	Not a primary focus
SCOPE	Global	Applies to AI systems placed on the EU market	Applies to Microsoft's AI development and deployment	Not directly applicable

Source: compiled by authors.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

5. Conclusion

The rise of AI presents both possibilities and difficulties. Despite its enormous potential to make our lives better, its capacity to alter reality calls for caution. To successfully navigate the complexities of a world increasingly shaped by Artificial Intelligence, it is imperative to develop strong regulatory frameworks, promote AI literacy, and foster critical thinking abilities along with technologically enhanced human abilities. In a digital world that is continuously changing, we can work to preserve a positive relationship with truth by exploiting AI's assistive qualities while minimizing any potential drawbacks.

AI possesses immense potential, but its misuse threatens our ability to discern truth. Combating misinformation is a manifold attempt that requires robust regulations, the development of reliable verification tools, and increased critical thinking capabilities among citizens. While the EU's AI regulation is a positive step, ongoing vigilance and adaptation will be necessary to navigate this increasingly hazy reality. Intelligence Augmentation (IA) that was developed in parallel with AI, displays an inexhaustible potential in securing the technology to remain in service of humans and democracy. However, while the conspiracy theories saturate the news media all over the world, Intelligence Augmentation finds much less media time. Nevertheless, using the technology to sustain and improve human detection capabilities could prove useful in tackling the perils of confabulation enhanced with Artificial Intelligence.

The disposition of individuals to retweet false information more than the truth drives its rapid dissemination, despite factors that would otherwise promote truthful content. Additionally, while discussions on misinformation often highlight the role of bots in spreading false news, the findings of this research suggest that human behavior plays a more significant role in the differential spread of falsehoods and truth than automated bots. Hence, the approach for strengthening truth in the age of AI and IA, should prioritize behavioral interventions (Vosoughi et al, 2018) that discourage the spread of misinformation as well as augmenting human critical thinking capacities psychologically and technologically, rather than solely targeting AI bot suppression. For a more resilient society formed by informed citizens where democracy can continue to thrive, the way Artificial Intelligence is deployed could be regulated and channeled towards fostering ethical and truthful human behavior. As part of such strategy, Intelligence Augmentation could bolster these objectives by arming the citizens with better tools for searching, claiming and detecting the truth.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

References

Oord, Aaron V.; Dieleman, Sander; Zen, Heiga; Simonyan, Karen; Vinyals, Oriol; Graves, Alex; Kalchbrenner, Nal; Senior, Andrew; Kavukcuoglu, Koray (2016). WaveNet: A Generative Model for Raw Audio. In *ArXiv*. https://arxiv.org/abs/1609.03499.

Adams, Dennis M. and Hamm, Mary (2001). *Literacy in a Multimedia Age*. MA: Christopher-Gordon Publishers.

AFP Indonesia (2023). Old Video of Mexico Explosion Falsely Shared as Israel-Hamas War Footage. In *AFP*. Retrieved 10 March 2024 at https://factcheck.afp.com/doc.afp.com.348Q4RW.

Aral, Sinan (2020). *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health - and How We Must Adapt.* New York: Crown Currency.

Artificial Intelligence Act (2023). Retrieved 10 March 2024 at https://bit.ly/3xBq7H0.

Barnidge, Matthew (2018). Social Affect and Political Disagreement on Social Media. In *Social Media • Society*, Vol.4, n3. DOI: https://doi.org/10.1177/2056305118797721.

Buchanan, Bruce (2006). A (Very) Brief History of Artificial Intelligence. In *Ai Mag*, n26: 53-60.

Cassauwers, Tom (2019). Can artificial intelligence help end fake news? In *Horizon: The EU Research and Innovation Magazine.* Retrieved 10 March 2024 at https://projects.research-and-innovation.ec.europa.eu/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news

Cole, Samantha (2017). AI-Assisted Fake Porn Is Here and We're All Fucked. In *Motherboard Tech by Vice.* Retrieved 10 March 2024 at https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Confessore, Nicholas (2018). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. In *New York Times.* Retrieved 6 April 2024 at https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Creeber, Glen and Martin, Royston (2008). *Digital Culture: Understanding New Media.* London: McGraw-Hill Education.

Danesi, Marcel (2017). Eco's Definition of Semiotics as the Discipline of Lying. In Thellefsen, Torkild and Sørensen, Bent (eds.) *Umberto Eco in His Own Words*. Berlin, Boston: De Gruyter Mouton. DOI: https://doi.org/10.1515/9781501507144-004.

Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai and Fei-Fei, Li (2009). ImageNet: A Large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 248-255. DOI: https://doi.org/10.1109/CVPR.2009.5206848.

Eco, Umberto (1976). *A Theory of Semiotics*. Bloomington: Indiana University Press.

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Ellery, Simon. (2023). Fake photos of Pope Francis in a puffer jacket go viral, highlighting the power and peril of AI. *CBS News*. Retrieved 10 March 2024 at at: https://www.cbsnews.com/news/pope-francis-puffer-jacket-fake-photos-deepfake-power-peril-of-ai/

Engelbart, Douglas (1962). Augmenting Human Intellect: A Conceptual Framework. In *Contract AF,* 49-1024.

Goodfellow, Ian J.; Mirza, Mehdi; Xu, Bing; Ozair, Sherjil; Courville, Aaron and Bengio, Yoshua (2014). Generative Adversarial Networks. In *ArXiv*. Retrieved 21 March 2024 at https://arxiv.org/abs/1406.2661

Gottfredson, Linda S. (1997). Mainstream Science on Intelligence: An Editorial with 52 Signatories, History and Bibliography. In *Intelligence,* n24: 13–23.

Groh, Matthew; Epstein, Ziv; Firestone, Chaz; and Picard, Rosalind (2021). Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds. In *Proceedings of the National Academy of Sciences*, Vol.119, n1. DOI: https://doi.org/10.1073/pnas.2110013119.

Groh, Matthew; Epstein, Ziv; Obradovich, Nick; Cebrian, Manuel; and Rahwan, Iyad (2021). Human Detection of Machine Manipulated Media. In *ArXiv*, *Communications of the ACM*, Vol.64, n10: 40-47.

Harford, Tim (2024). It's Only a Matter of Time Before Disinformation Leads to Disaster: Fakes, Forgeries and the Meaning of Meaning in Our Post-truth Era. In *Financial Times Magazine*. Retrieved 10 March 2024 at https://www.ft.com/content/0afb2e58-c7e2-4194-a6e0-927afe0c3555

Harize, Ouissal (2023). AI Generated Image Shared as Greek Rescuer Holding a Turkish Child. In *Misbar*. Retrieved 10 March 2024 at https://misbar.com/en/factcheck/2023/02/10/ai-generated-image-shared-as-greek-rescuer-holding-a-turkish-child

Hassani, Hossein; Silva, Emmanuel Sirimal; Unger, Stephane; TajMazinani, Maedeh; Mac Feely, Stephen (2020). Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? In *AI,* Vol.1, n2:143-155. DOI: https://doi.org/10.3390/ai1020008.

Huang, Guanxiong; Wang, Sai (2023). Is Artificial Intelligence More Persuasive than Humans? A Meta-analysis. In *Journal of Communication*, Vol.73, n6: 552-562. DOI: https://doi.org/10.1093/joc/jqad024.

Jones, Dorian (2024). Deepfake Videos Used in Local Elections in Turkey as Erdogan Battles for Istanbul. In *RFI.* Retrieved 10 March 2024 at https://www.rfi.fr/en/podcasts/international-report/20240316-deepfake-videos-used-in-local-elections-in-turkey-as-erdogan-battles-for-istanbul

DOI: http://dx.doi.org/10.6035/adcomunica.8020

Joy, Bill (2000). Why the Future Doesn't Need Us. *Wired Magazine*. Retrieved 10 March 2024 at https://www.wired.com/2000/04/joy-2/.

Karras, Tero; Laine, Samuli; Aittala, Miika; Hellsten, Janne; Lehtinen, Jaakko; and Timo, Aila (2019). Analyzing and Improving the Image Quality of StyleGAN. In *ArXiv*. Retrieved 8 April 2024 at https://arxiv.org/abs/1912.04958.

Kaczynski, Theodore (1995). The Unabomber Manifesto1. In *Industrial Society and its Future*. Retrieved 10 April 2024 at ftp.ai.mit.edu/pub/users/misc/unabomber

Kim, Alex and Yoon, Sangwon (2023). Detecting Rumor Veracity with Only Textual Information by Double-Channel Structure, In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, 35-44, DOI: https://doi.org/10.48550/arXiv.2312.03195

Krizhevsky, Alex; Sutskever, Ilya and Hinton, Geoffrey E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. In *Communications of the ACM*, Vol.60, n6: 84-90. DOI: https://doi.org/10.1145/3065386

Kurzweil, Ray (2001). InResponse to. *The Kurzweil Library and collections: Essays.*

Retrieved 10 April 2024 at https://www.thekurzweillibrary.com/in-response-to

Legg, Shane and Hutter, Marcus (2007). A Collection of Definitions of Intelligence. In Goertzel, Ben and Wang, Pei (Eds.) *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithm.* Amsterdam: IOS Press Ebooks.

McCarthy, Bill (2023). Ron DeSantis ad uses AI-generated photos of Trump, Fauci. *AFP*. Retrieved 10 March 2024 at https://factcheck.afp.com/doc.afp.com.33H928Z

McCarthy, John; Minsky, Marvin L; Rochester, Nathaniel; and Shannon, Claude E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. In *AI Magazine,* Vol.27, n4: 12.

McLuhan, Marshall (1951). *The Mechanical Bride: Folklore of Industrial Man.* New York: Vanguard Press.

Mercier, Hugo (2017). How Gullible are We? A Review of the Evidence from Psychology and Social Science. In *Review of General Psychology*, Vol.21, n2: 103-122. DOI: https://doi.org/10.1037/gpr0000111.

Pajnik, Mojca (2024). Professionalizing Emotions as Reflective Engagement in Emerging Forms of Journalism. In *Journalism Studies*, Vol.25, n2: 181–198. DOI: https://doi.org/10.1080/1461670X.2023.2289920.

Parkin, Simon (2019). The Rise of the Deepfake and the Threat to Democracy. In *The Guardian*. Retrieved 10 March 2024 at https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy

Parks, Perry (2021). Joy is a News Value. In *Journalism Studies*, Vol.22, n6, 820–838. DOI: https://doi.org/10.1080/1461670X.2020.1807395.

Paul, Christopher and Matthews, Miriam (2016). The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It. *RAND Corporation*. Retrieved 10 March 2024 at https://www.rand.org/pubs/perspectives/PE198.html

Purbrick, Louise (1993). The Dream Machine: Charles Babbage and His Imaginary Computers. In *Journal of Design History*, Vol.6, n1: 9-23.

Richards, Barry and Rees, Gavin (2011). The Management of Emotion in British Journalism. In *Media, Culture & Society*, Vol.33, n6: 851–867. DOI: https://doi.org/10.1177/0163443711411005.

Rossini, Patricia; Hemsley, Jeff; Tanupabrungsun, Sikana, Zhang, Feifei, and Stromer-Galley, Jennifer (2018). Social Media, Opinion Polls, and the Use of Persuasive Messages During the 2016 US Election Primaries. *Social Media • Society*, Vol.4, n3. DOI: https://doi.org/10.1177/2056305118784774

Safi, Michael; Atack, Alex and Kelly, Joshua (2024). Revealed: The Names Linked to ClothOff, the Deepfake Pornography App. In *The Guardian*. Retrieved 10 March 2024 at https://www.theguardian.com/technology/2024/feb/29/clothoff-deepfake-ai-pornography-app-names-linked-revealed

Schudson, Michael (2001). The Objectivity Norm in American Journalism. In *Journalism,* Vol.2, n2: 149-170. DOI: https://doi.org/10.1177/146488490100200201.

Spitale, Samuel C. (2022). *How to Win the War on Truth? An Illustrated Guide to How Mistruths Are Sold, Why They Stick and How to Reclaim Reality.* Philadelphia: Quirk.
Spring, Marianna (2024). Trump supporters target black voters with faked AI images. In *BBC Panorama and Americast.* Retrieved 16 March 2024 at https://www.bbc.com/news/world-us-canada-68440150

Tarchi, Pietro; Lanini, Maria Chiara; Frassineti, Lorenzo; and Lanatà, Antonio (2023). Real and Deepfake Face Recognition: An EEG Study on Cognitive and Emotive Implications. *Brain Sciences*, Vol.13, n9: 1233. DOI: https://doi.org/10.3390/brainsci13091233.

Tegmark, Max (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Alfred A. Knopf.

Toffler, Alvin (1981). *The Third Wave*. New York: Bantam Books.

Trusted Web Foundation (2021). Manifest. Retrieved 14 March 2024 at https://thetrustedweb.org/manifest/

Ulmer, Alexandra and Tong, Anna (2023). With Apparently Fake Photos, DeSantis Raises AI Ante. Reuters. Retrieved 10 March 2024 at https://www.reuters.com/

DOI: http://dx.doi.org/10.6035/adcomunica.8020

world/us/is-trump-kissing-fauci-with-apparently-fake-photos-desantis-raises-ai-ante-2023-06-08/

UNESCO (2022). Freedom of Expression, Media and Information Literacy and Digital Competencies to support peace and human rights. Retrieved 10 March 2024 at https://unesdoc.unesco.org/ark:/48223/pf0000381532

Van Der Lans, Sebastiaan (2021). 13 AI-Powered Tools for Fighting Fake News. *Trusted Web Foundation*. Retrieved 10 March 2024 at https://thetrustedweb.org/ai-powered-tools-for-fighting-fake-news/

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; and Polosukhin, Illia (2017). Attention Is All You Need. *ArXiv*, Retrieved 10 March 2024 at: 14.10.2024. /abs/1706.03762

Vosoughi, Soroush; Roy, Deb and Aral, Sinan (2018). The Spread of True and False News Online. *Science,* n359, 1146-1151. DOI: https://doi.org/10.1126/science.aap9559.

DOI: http://dx.doi.org/10.6035/adcomunica.8020