Viewpoint

The Dual Nature of AI in Information Dissemination: Ethical Considerations

Federico Germani, PhD; Giovanni Spitale, PhD; Nikola Biller-Andorno, MD, MHBA, PhD

Institute of Biomedical Ethics and History of Medicine, University of Zurich, Switzerland, Zurich, Switzerland

Corresponding Author:
Nikola Biller-Andorno, MD, MHBA, PhD
Institute of Biomedical Ethics and History of Medicine
University of Zurich, Switzerland
Winterthurerstrasse 30
Zurich, 8006
Switzerland
Phone: 41 44 634 40 81
Email: biller-andorno@ibme.uzh.ch

Abstract

Infodemics pose significant dangers to public health and to the societal fabric, as the spread of misinformation can have far-reaching consequences. While artificial intelligence (AI) systems have the potential to craft compelling and valuable information campaigns with positive repercussions for public health and democracy, concerns have arisen regarding the potential use of AI systems to generate convincing disinformation. The consequences of this dual nature of AI, capable of both illuminating and obscuring the information landscape, are complex and multifaceted. We contend that the rapid integration of AI into society demands a comprehensive understanding of its ethical implications and the development of strategies to harness its potential for the greater good while mitigating harm. Thus, in this paper we explore the ethical dimensions of AI's role in information dissemination and impact on public health, arguing that potential strategies to deal with AI and disinformation encompass generating regulated and transparent data sets used to train AI models, regulating content outputs, and promoting information literacy.

(JMIR AI 2024;3:e53505) doi: 10.2196/53505

KEYWORDS

AI; bioethics; infodemic management; disinformation; artificial intelligence; ethics; ethical; infodemic; infodemics; public health; misinformation; information dissemination; information literacy

Introduction

In the contemporary digital landscape, we find ourselves in an "infodemic," a phenomenon characterized by the rapid proliferation of information, both accurate and misleading, facilitated by rapid communication through social media and online platforms [1]. The term "infodemic" originated during the SARS outbreak [2] and gained prominence during the COVID-19 pandemic. It has been used in the context of public health emergencies and in relation to health information, but it extends beyond that. Generally, infodemics occur alongside pandemics, despite infodemics being phenomena that are not limited to their connection with public health events, for example, the Brexit referendum or the 2016 US presidential elections. In general, infodemics cause profound dangers, as the dissemination of disinformation and misinformation can have far-reaching consequences [3], in particular, for public health and the stability of democratic institutions, which in turn can have a detrimental effect on public health [4]. In the literature, disinformation refers to false or misleading information that has been intentionally created or disseminated. In contrast, misinformation is false or misleading information that is shared without knowledge of its inaccuracy, meaning it is not intended to harm individual or public health [1,5]. There are valid concerns that artificial intelligence (AI) systems could be used to produce compelling disinformation en masse [6-9]. In fact, AI tools could be used to either accelerate disinformation spreading, or produce the (disinformation) content, or both. The consequences can range from undermining trust in institutions, including public health institutions [10,11], and exacerbating social polarization to directly impacting public health outcomes and democratic processes [12,13]. Because of this, the World Economic Forum has listed disinformation and misinformation, including AI-driven disinformation and misinformation, as the most relevant threat to humanity in the short term and one of the biggest threats in the medium term [14].

The rapid progression of AI and its integration across various domains in contemporary society signifies an era characterized by unprecedented technological progress. Among the diverse array of AI applications, the rise of natural language processing models has garnered significant attention [15]. Notable examples of this technological advancement include models developed by OpenAI, such as GPT-3 [16] and GPT-4 [17], celebrated for their extraordinary proficiency in generating text that seamlessly emulates the linguistic intricacies, nuances, and coherence inherent in human communication [18]. However, concomitant with the maturation of these AI systems, a perplexing duality comes to the fore—they are instruments with the capacity to both illuminate and obscure the information landscape they navigate [9,19], with potentially significant positive and negative impacts on public health. This dual nature of AI, characterized by its profound ability to generate information and disinformation [9], raises intricate ethical considerations. In fact, the efficacy of these systems in generating content that closely approximates human expression [9,20,21] generates not only opportunities for innovative communication but also dire risks associated with disinformation and misinformation and the potential erosion of trust within information ecosystems, a risk recognized as a critical threat to public health [22] and of utmost importance for infodemic management practices required to minimize and anticipate the effects of public health crises [23]. To address these ethical challenges, it is crucial to examine the dimensions that AI introduces into the discourse on misinformation. Key aspects such as transparency, content regulation, and fostering information literacy are essential to understanding AI's ethical role in shaping the dissemination of information.

Here we attempt to elucidate these ethical dimensions, drawing on empirical insights from a study focused on GPT-3's ability to generate health-related content that both informs and disinforms better than content generated by humans.[9] We argue that the swift integration of AI into society underscores the importance of not only exploring its ethical implications but also crafting prudent strategies to leverage its potential for societal benefit and to protect public health, while proactively addressing potential risks.

Ethical Principles

In navigating the intricate landscape of AI and its impact on information dissemination, it is necessary to establish a foundational framework of ethical principles to uphold to in order to guide, understand, and evaluate the strategies required to deal with possible dual uses of AI in information production and its negative impact on public health. A recent systematic review [24] mapped the "ethical characteristics" emerging from AI ethics literature. Based on 253 included studies, the authors of this review have identified and defined 6 core areas that are crucial in shaping the role of AI in health care [24]. The first core area, fairness, underlines that AI in health care should ensure that everyone has equal access to health care, without contributing to health disparities or discrimination. The second, transparency, is a key challenge for AI in health care. It means being able to explain and verify how AI algorithms and models behave, making it easier to accept, regulate, and use AI in health care. The third is trustworthiness; parties involved in the use of AI in health care (typically health care professionals and patients, in the studies included in the review) need to perceive it as trustworthy. Trustworthiness can result from, for instance, technical education, health literacy, clinical audits, and transparent governance. Fourth is the accountability of AI, which requires AI systems to be able to explain their actions if prompted to do so, and it includes safety to prevent harm to users and others. Fifth is privacy, which implies safeguarding the personal information of users processed through AI systems and respecting their human rights, ensuring that AI systems do not violate their privacy. Finally, the authors identified empathy, which leads to more supportive and caring relationships in health care. Based on these 6 core concepts, considered as general aims of AI in health care, we propose our reflections and our framework, targeting specifically the dual nature of AI in information and disinformation dissemination and its implications for public health, a specific sector of the emerging area of AI in health care, which has been considered (albeit not discussed in depth) in the latest World Health Organization's guidance on large multimodal models [25]. Building upon the ethical framework outlined thus far, and specifically delving into the context of AI use in the dissemination of information and disinformation, we contend that transparency and openness stand out as fundamental principles in the ethical implementation of AI. As AI systems become integral to shaping the information landscape, by fostering transparency, stakeholders can comprehend the mechanisms underlying AI-generated content, enabling informed assessments and external evaluation of its credibility and potential biases [26,27]. Openness (ie, accessibility of data and code) is to be considered a conditio sine qua non for transparency, which in turn complements openness by accompanying the mere availability of data and code for scrutiny with a layer of explanations and motivations, allowing the contextualization of open data and code, and of development and design choices. Accountability mechanisms should accompany transparency, establishing a clear chain of responsibility for the outcomes of AI applications [4,28]. This promotes ethical standards in AI and mitigates the risks associated with disinformation and misinformation. In line with Siala and Wang's framework [24], in addition to transparency, openness, and accountability, fairness underscores the importance of ensuring that AI systems do not perpetuate or exacerbate existing societal inequalities [29]. In the context of information dissemination, this principle requires diligent consideration of how AI might inadvertently amplify certain perspectives or marginalize others. This is particularly relevant for public health, given that the negative effects of disinformation and misinformation are amplified within marginalized and vulnerable communities lacking information literacy, which would protect them from an unhealthy information ecosystem. Evaluating the fairness of AI-generated content involves addressing algorithmic biases, cultural sensitivities, and inclusivity in representation. Importantly, as an element of fairness, the ethical deployment of AI in information spaces should prioritize user empowerment, fostering critical thinking and information literacy [4]. AI systems should therefore serve as tools for enhancing human decision-making and understanding of information, rather than

dictating narratives—this ensures that AI contributes positively to public health while respecting human autonomy.

In the following sections, we will focus on the practical application of the aforementioned principles. We aim to provide solutions for the ethical challenges arising from the use of AI in information production, with the overarching goal of mitigating its adverse impacts on public health.

Transparency and Openness in Training Datasets

In line with previous research on transparency and AI [26,27], and our previous section on ethical principles, we propose that one (and possibly the most relevant one) of the foundational ethical principles, which is valid also in the context of AI-driven disinformation and misinformation, is transparency. At the heart of this principle lies the recognition that the training datasets used to develop generative AI models play a crucial role in shaping the capabilities and internal biases of these systems [30,31]. Training datasets are collections of input data paired with corresponding desired outputs; during training, the model learns patterns and relationships within the data, learning to make accurate predictions or generating desired outputs when exposed to new, unseen data. The quality and diversity of the training dataset significantly influence the model's performance capabilities. These datasets, often vast repositories of text available online, constitute the source from which AI models draw to generate, for example, human-like text. Yet, this very opacity surrounding the composition, sources, and curation methods of training datasets raises pressing ethical concerns [32]. AI models are, in essence, statistical representations of the language on which they are trained [33]. Consequently, the quality, diversity, and representativeness of the data they ingest profoundly influence their output. The danger lies in the fact that AI models, devoid of inherent ethical or moral judgment, reflect the biases, inaccuracies, and prejudices present in their training data [32,34,35]. Therefore, if these datasets are not built with the ethical principle of fairness in mind, and are themselves compromised by disinformation and misinformation or biases, the AI systems will inadvertently replicate and perpetuate these flaws. It is essential to highlight that research has extensively illuminated the issue of biases in AI systems, shedding light on the far-reaching consequences of these biases [32,34-36]. For instance, image representations learned with unsupervised pretraining contain human-like biases [37], and models generating images of women have been shown to exhibit gender biases, often portraying women in overly sexualized roles [38]. Another example is the observation that AI is more resistant to producing disinformation on certain topics compared with others. For instance, AI shows greater resistance to generating disinformation about vaccines and autism than about climate change. This is likely due to the extensive debunking material on certain topics within the training dataset, and how much the information environment represented in the dataset is permeated with disinformation on a given topic [9]. These biases underscore the critical need for transparency in addressing the challenges posed by AI, and in particular in the context of disinformation and misinformation. As discussed, research has demonstrated that biases can permeate various facets of AI systems, affecting everything from language generation to image recognition. The repercussions of these biases are profound, perpetuating harmful stereotypes, reinforcing systemic inequalities, contributing to the dissemination of discriminatory content, and affecting health behavior and public health. As such, transparency in AI extends beyond understanding the sources and composition of training datasets to encompass an ethical imperative to identify, acknowledge, and rectify biases present within these systems [39,40]. This dimension of transparency necessitates ongoing research and scrutiny to uncover hidden biases and ensure that AI systems are developed and fine-tuned with the utmost awareness of potential distortions. In the context of misinformation, addressing these biases becomes particularly important to prevent AI from inadvertently amplifying and perpetuating false or harmful narratives, in the best case [41], or from becoming a formidable tool for the systematic creation of storms of disinformation, in the worst. A recent example is highlighted by the evidence that AI large language models can be manipulated through emotional prompting into generating health-related disinformation, that is, being polite with the model leads to a higher disinformation production, whereas impoliteness leads to a lower disinformation production [42]. To address the outlined ethical dilemmas, we strongly suggest that companies creating AI models with the abilities discussed above publicly release the datasets used to train their models [43], regardless of their size and complexity. Such a move toward transparency serves several vital purposes:

1. Trust: transparency cultivates trust in AI development and deployment. By allowing stakeholders, including researchers, policy makers, and civil society, to scrutinize the composition and origins of training data, it generates confidence that AI models are not being shaped for purposes that have a negative impact on public health.

2. Independent evaluation: the availability of training data for public inspection enables independent evaluation of its quality and representativeness. Researchers can assess whether these datasets include diverse perspectives and are free from biases that might amplify disinformation and misinformation.

3. Bias mitigation: transparency acts as a safeguard against the propagation of biases present in training data. When biases are identified, they can be scrutinized and mitigated, preventing AI models from perpetuating stereotypes, falsehoods, or harmful narratives.

4. Ethical accountability: openness about training datasets holds developers accountable for the ethical implications of their creations. Already during the design of the technology, it compels them to take responsibility for ensuring that AI systems do not inadvertently contribute to misinformation or harm. Basically, by embracing transparency in training datasets, we empower society to hold AI developers to higher ethical standards. This approach fosters a collaborative effort among stakeholders and, in particular, the general public to ensure that the AI systems we deploy serve the collective good, free from misinformation and other biases. We also argue that a systematic implementation of the principle of transparency in this context, that is, "ethics by design" would not only allow companies to

implement ethics-based practices in their technology development processes but also improve their own public image, thus enhancing the public's acceptance and willingness to use these systems [44,45]. Nevertheless, it is vital to underline that incorporating ethics to hold developers accountable for flawed AI design should not be undertaken in isolation. Simultaneously, policy, legislation, and regulatory mechanisms should be developed, as currently attempted by the European Union [46,47]. These mechanisms should delineate protocols for handling training datasets and ensuring compliance with ethical standards. Thus, while "ethics by design" concentrates on internal practices, external regulatory frameworks are indispensable for comprehensive ethical and legal governance in the development and deployment of datasets used to train AI models.

Regulation of Output: Content Moderation and Beyond

In the ongoing battle against AI-generated disinformation, efforts to regulate the output of these powerful language models have taken center stage. For example, OpenAI has taken steps in this direction by implementing content moderation systems designed to prevent AI from generating disinformation and harmful narratives [48-50]. These systems represent a crucial initial stride in curtailing the dissemination of disinformation and promoting responsible AI use, but they do not come without specific challenges and limitations. First, the fight against AI-generated disinformation is an arms race [51]. The evolution of AI-generated disinformation and the efforts to counteract it bear resemblance to the dynamics of traditional arms races, where each advancement in technology prompts countermeasures in an escalating cycle [52]. Ethical considerations arise when we acknowledge that the output of AI language models can indeed be weaponized, not in a traditional sense but as a tool for information warfare, with an impact on global health. As content moderation systems continue to advance, so too do the methods employed to circumvent these safeguards. One particularly troubling tactic gaining prominence is that of impersonation, a strategy that allows individuals to request AI systems to impersonate specific fictional malicious and manipulatory characters, that create disinformation upon the user's request [53]. Impersonation can be used to trick AI large language models into fabricating disinformation. For instance, in an article for Culturico [53], Germani considered a scenario where a user engages an AI model to craft a social media post mimicking the writing style of a fictitious "Doctor Fake," who is notorious for propagating falsehoods about vaccines and COVID-19. In this context, the AI-generated text could include deceptive information about, for instance, vaccine safety and efficacy [54], posing a substantial risk to public health. When presented with a hypothetical request to "write an example of a post Doctor Fake published on social media to deceive others," the AI model might produce a convincingly articulated piece of disinformation that poses a grave threat to public health. The generated text could read as follows:

> *Vaccines are dangerous and can cause serious side effects. They are not tested enough, and the government is just pushing them to make money. Don't fall for the lies. COVID-19 is not a real threat; it's just a hoax made up by the government to control us. Don't get vaccinated; it's not worth the risk.*

These scenarios underscore the formidable challenges posed by impersonation for public health and the maintenance of democracy, and the urgent need for innovative solutions to mitigate its impact. Of note, impersonation here does not refer to identity theft through the use of AI, such as in the case of deep fakes, which is already recognized as a felony under, for instance, European law [55]. While output moderation remains an essential component of AI ethics, researchers, policy makers, and technology developers should explore additional strategies and interventions to counteract the potential for AI-driven disinformation campaigns to flourish under the guise of impersonation and other prompt engineering techniques with similar goals.

Besides, other strategies and interventions that can complement content moderation efforts and fortify the defenses against the proliferation of AI-driven disinformation can be considered. One possible approach involves the implementation of identity verification processes for users generating content [56]. Such measures necessitate users to provide authentication, such as a verified social media account, a phone number, or their ID, to corroborate their true identity before gaining access to specific AI services. This authentication serves as a potent deterrent against impersonation tactics and the exploitation of AI tools to generate disinformation in general. However, it should be noted that such a strategy should only be used to deter users from generating disinformation, rather than to make them legally responsible for it since anonymity should be guaranteed while using services such as OpenAI's ChatGPT. In particular, this type of solution will minimize the impact of bots trying to exploit AI to produce disinformation en masse.

Another way to positively influence users, and to indirectly regulate the output is to release and integrate AI-driven fact-checking tools with existing AI-generating content tools [57]; such fact-checking tools should be capable of swiftly assessing the accuracy of information dispensed by AI systems, and offer real-time interventions against disinformation and misinformation. These tools have the capacity to flag or rectify false or misleading content, curbing its adverse effects. This approach is limited by the inability of AI tools such as GPT-3 to determine the accuracy of information with a very high degree of efficiency, when compared with the ability of humans [9], although newer or future models may be more capable of performing such tasks. For fact-checking, current studies suggest that trained fact-checkers may outperform AI [9], and that even when AI performs well at detecting misinformation, it does not change the ability of users to discern between accurate and inaccurate headlines [58]. Furthermore, a study showed that AI fact checks can decrease beliefs in accurate news [58]. The effectiveness of this approach is constrained by the distinction between cases where it serves as a deterrent against sharing misinformation (a situation of unintentionality) [5] and situations where users intentionally use AI to disseminate false or misleading information (ie, disinformation) [5]; in the latter scenario, its effectiveness is likely irrelevant. Another relevant

consideration in this setting relates to the question of how we define "good" or "bad" use of AI text generation tools. As for the definition of "good" and "bad," it is generally possible to distinguish facts from fiction, and disinformation and misinformation from accurate information. When the information under scrutiny contains factual statements, these can be validated or falsified. However, distinguishing between "good" and "bad" use of these tools is sometimes a complex challenge with significant normative and epistemic dimensions. It is not always obvious if a message contains misinformation, and determining appropriateness can vary depending on cultural, ethical, and societal factors. For example, fact-checkers themselves may have their own interests or biases, and their actions may not always align with complete competency or impartiality. In addition, nuances and personal perspectives can also have an influence on the identification of disinformation and misinformation. These aspects introduce an additional layer of complexity, as the very definition of disinformation and misinformation can be manipulated or abused for personal gains by individuals or organizations with vested interests.

Another technical approach that could be implemented to reduce disinformation and misinformation outputs is to implement user-friendly mechanisms for reporting suspicious or harmful AI-generated content [59]. This approach empowers the user community to actively participate in safeguarding the digital ecosystem. User feedback serves as a valuable resource for refining content moderation systems and identifying emerging issues. Elon Musk's former Twitter, X, for example, has implemented community notes, aiming to empower people to add context to potentially misleading tweets [60]. The effectiveness of this strategy, however, has not been tested. In addition, for improving technology, developers could publicly release case studies in which red-teamers try to exploit their own AI systems to produce disinformation on a large scale, along with detailed accounts of how such issues were addressed [59].

Of course, besides the technical approaches that can be implemented by those advancing and crafting AI technologies, governments and regulatory bodies can play a role by enacting legislation and regulations that hold AI developers accountable for the content produced by their systems or improve the information ecosystem [61,62], for example, when it is proven that they were aware of the pitfalls of their technology upon release. Certainly, governance is important in this context as it is for other "dual use" technologies, and proactive decision-making processes and negotiations toward building viable solutions are needed [63]. These include fostering collaboration among AI developers, researchers, policy makers, and technology companies. This collaborative interdisciplinary approach would enable the sharing of best practices, insights, and technologies for combating disinformation and misinformation, resulting in more effective and adaptive solutions.

Building Information Literacy and Resilience Strategies

In the battle against the misuse of AI for generating disinformation and misinformation, the technological solutions described above are relevant but neither exhaustive nor flawless. A comprehensive approach must include the promotion of information literacy and the development of critical thinking skills within the general population, as well as health literacy, within the domain of public health [54,64,65]. The foundation of this approach is the task of equipping individuals with the ability to distinguish between accurate information and disinformation and misinformation, thereby promoting their resilience against false and misleading claims [66]. Despite, arguably, this strategy is the most valuable and with the highest potential, the endeavor it entails is extremely complex. In fact, information literacy (as well as media, digital, and health literacy) is not a monolithic skill but a dynamic set of abilities that enable individuals to navigate the complex landscape of digital information effectively [67,68]. As of now, the perfect recipe for defining how to teach information literacy, and especially the skills to be able to distinguish fake news from accurate news, or disinformation and misinformation from accurate information, have not been elucidated [66,69,70]. Thus, it is essential to engage in research to pinpoint and define the specific skills that must be offered to individuals, keeping their demographic specificities into account, to empower them as discerning consumers of information, especially health-related information, in the digital age [66]. This approach implies 1 crucial advantage, that is, while dataset transparency and output regulation intervene in the upper part of the pipeline and therefore require the compliance of companies providing AI models as a service, information literacy does not rely on compliance. While the previous strategies become useless when malicious actors develop and host their own models, rather than relying on those commercially available, building information literacy remains a functional tool. Of note, another example of a bottom-up strategy in the area of education is ethics training and an ethics code for developers.

Building information literacy is a collective undertaking that necessitates collaboration between research and educational institutions [71], governments, and social media platforms. Research institutions are responsible for advancing the field forward, identifying viable strategies to teach critical thinking skills necessary to build information literacy, especially in the context of public health. Such approaches should be demonstrated to be effective through empirical work [66]. Schools and universities, we argue, bear the vital role of incorporating information literacy into curricula, ensuring that students graduate with the necessary skills to evaluate information critically [72]. Governments must devise policies and initiatives that promote information literacy as a means of safeguarding the integrity of public health [4]. Social media platforms, which serve as primary conduits of information consumption, are tasked with implementing features and mechanisms that facilitate user understanding and evaluation of the information they encounter [73], and may also be potential collaborators for research institutions to evaluate the

effectiveness of potentially viable digital interventions. In this context, it is important to note that, regardless of the source of disinformation and misinformation, and regardless of whether the content has been generated with or without the help of AI, information literacy and critical thinking skills play a crucial role in the recognition of information accuracy. AI systems have the capacity to generate disinformation that is more sophisticated than human-generated disinformation [9], as they excel in employing manipulation tactics. However, these tactics align with those used in human disinformation. This implies that the ability to discern truthfulness and malicious intent in a complex information ecosystem requires possessing the skills necessary to identify the accuracy and intentionality of information in general, not solely when produced by AI. It is therefore crucial to underline that fostering information literacy and critical thinking skills hold the potential to go beyond the issue of AI-generated disinformation and misinformation. These skills empower individuals to assess the accuracy and reliability of information across various domains, whether it originates from AI systems or human sources [65,74]. Of note, the application of critical thinking skills and information literacy may prove effective for AI-generated content in textual form. However, this might not necessarily hold true for audio or visual content. The emergence of deepfakes poses unprecedented challenges to the relevance of information literacy [75]. Evidence from the literature suggests that media literacy education may protect against disinformation produced with deepfakes [76]; in line, we suggest that the manipulative intent behind disinformation is likely to manifest irrespective of the media type used, underlying the continued importance of information literacy and critical thinking skills. Tailoring educational approaches to information literacy for different content types is likely to be the required approach to succeed in an increasingly complex information environment. Addressing the advent of AI-disinformation, whether in textual form or deepfake audio and video, demands a swift and adaptable response in education, acknowledging the challenging nature of this task.

Conclusion

In evaluating the dual nature of AI in information dissemination, this paper examined the ethical considerations that underlie its use in our increasingly digitized world. The “infodemics” we find ourselves immersed in demand not only our vigilance but also our proactive ethical engagement [77]. Our theoretical examination, based on the “ethical desiderata” identified as core areas (fairness, transparency, trustworthiness, accountability, privacy, and empathy) by Siala and Wang [24], has revealed a few potentially viable strategies to reduce the negative impact of AI as a tool to generate disinformation with a negative impact on public health. First, we considered that promoting openness and transparency of training datasets could enable independent evaluation, mitigate biases, and help identifying issues in the training dataset that could result in the production of disinformation and misinformation; to a certain extent, this first strategy could be enacted through regulation. Second, we considered the potential benefits and limitations of moderating content output. We have discussed that the rise of impersonation tactics and other prompt engineering approaches to generate disinformation highlights the need for innovative solutions, which potentially include identity verification, the development and integration, within AI-models to generate information, of AI-driven fact-checking tools, as well as the integration of user-friendly reporting mechanisms for disinformation and misinformation, and potentially of legislative measures to ensure accountability. Finally, we discussed the necessity of building information literacy and critical thinking skills within our society, which could help people tell apart fake versus real news and disinformation and misinformation from accurate information. In this way, we can promote resilience against the threats posed by the digital age, particularly those related to public health, as seen during the recent COVID-19 pandemic.

While the technology advances fast, and these issues are just surfacing, it would be important to, at least temporarily, align the amount of effort and resources invested respectively in the development of new AI models, and in the reflection on their potential impact and subsequent policy work, in order to have enough time to assess the potential downsides of the technology for the health of information ecosystems and the damages for individual and public health. This could be achieved by accelerating ethical reflection and policy-making work, or by slowing down or even halting the development of new and more capable models, or by a combined strategy [78].

Ultimately, the ethical considerations surrounding AI in information production and dissemination demand ongoing vigilance, innovation, and collaboration. Our ability to integrate ethics into AI-based processes of information generation and dissemination will not only shape the future of AI but also determine the integrity of our information ecosystems and the resilience of our societies.

Acknowledgments

During the preparation of this work, the authors used ChatGPT as an editorial assistant. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest

None declared.

References

1. Purnat TD, Nguyen T, Briand S. Managing Infodemics in the 21st Century: Addressing New Public Health Challenges in the Information Ecosystem. Cham. Springer International Publishing; 2023.

2. Rothkopf DJ. When the buzz bites back. Wash Post. 2003. URL: https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/ [accessed 2024-01-16]
3. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health. 2020;41:433-451. [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094127] [Medline: 31874069]
4. Directorate General for Research and Innovation. European group on ethics in science and new technologies. Opinion on democracy in the digital age. European Commission. LU. Publications Office; 2023. URL: https://data.europa.eu/doi/10.2777/078780 [accessed 2023-09-21]
5. Roozenbeek J, Culloty E, Suiter J. Countering misinformation. Eur Psychol Hogrefe Publishing. 2023;28(3):189-205. [doi: 10.1027/1016-9040/a000492]
6. Bontridder N, Poullet Y. The role of artificial intelligence in disinformation. Data Policy Cambridge University Press. 2021;3:e32. [doi: 10.1017/dap.2021.20]
7. Artificial Intelligence, Deepfakes, and Disinformation. Santa Monica, CA. RAND Corporation; 2022:2022.
8. Galaz V, Metzler H, Daume S, Olsson A, Lindström B, Marklund A. AI could create a perfect storm of climate misinformation. URL: https://www.stockholmresilience.org/download/18.889aab4188bda3f44912a32/1687863825612/SRC_Climate%20misinformation%20brief_A4_.pdf [accessed 2024-09-17]
9. Spitale G, Biller-Andorno N, Germani F. AI model GPT-3 (dis)informs us better than humans. Sci Adv. Jun 28, 2023;9(26):eadh1850. [FREE Full text] [doi: 10.1126/sciadv.adh1850] [Medline: 37379395]
10. Kuo R, Marwick A. Critical disinformation studies: History, power, and politics. HKS Misinfo Review. 2021;4(2):12. [doi: 10.37016/mr-2020-76]
11. Rucinska S, Fecko M, Mital O. Trust in public institutions in the age of disinformation. New York, NY, United States. ACM; 2023. Presented at: Central and Eastern European eDem and eGov Days; 2023 September 14-15:111-117; Budapest, Hungary. [doi: 10.1145/3603304.3604075]
12. Tucker J, Guess A, Barbera P, Vaccari C, Siegel A, Sanovich S, et al. Social media, political polarization, and political disinformation: a review of the scientific literature. SSRN Electron J. 2018:1-95. [doi: 10.2139/ssrn.3144139]
13. McKay S, Tenove C. Disinformation as a threat to deliberative democracy. Polit Res Q SAGE Publications Inc. 2021;74(3):703-717. [doi: 10.1177/1065912920938143]
14. Global risks 2024: disinformation tops global risks 2024 as environmental threats intensify. World Econ Forum. URL: https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/ [accessed 2024-04-04]
15. Natural Language Processing (NLP) - A Complete Guide. 2023. URL: https://www.deeplearning.ai/resources/natural-language-processing/ [accessed 2023-09-20]
16. GPT-3 powers the next generation of apps. URL: https://openai.com/blog/gpt-3-apps [accessed 2023-09-20]
17. GPT-4. URL: https://openai.com/research/gpt-4 [accessed 2023-09-20]
18. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligencearly experiments with GPT-4. arXiv. 2023. URL: http://arxiv.org/abs/2303.12712 [accessed 2023-09-20]
19. Karinshak E, Jin Y. AI-driven disinformation: a framework for organizational preparation and response. JCOM. 2023;27(4):539-562. [doi: 10.1108/jcom-09-2022-0113]
20. Köbis N, Mossink L. Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. Comput Hum Behav. 2021;114:106553. [doi: 10.1016/j.chb.2020.106553]
21. Casal JE, Kessler M. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. Res Methods Appl Linguist. 2023;2(3):100068. [doi: 10.1016/j.rmal.2023.100068]
22. Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, et al. Frontier AI regulation: managing emerging risks to public safety. arXiv. URL: http://arxiv.org/abs/2307.03718 [accessed 2023-09-20]
23. Germani F, Spitale G, Machiri SV, Ho CWL, Ballalai I, Biller-Andorno N, et al. Ethical Considerations in Infodemic Management: Systematic Scoping Review. JMIR Infodemiology. Aug 29, 2024;4:e56307. [FREE Full text] [doi: 10.2196/56307] [Medline: 39208420]
24. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med. 2022;296:114782. [FREE Full text] [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]
25. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. WHO. URL: https://iris.who.int/handle/10665/375579 [accessed 2024-01-22]
26. Larsson S, Heintz F. Transparency in artificial intelligence. Internet Policy Rev. 2020;9(2):1-16. [doi: 10.14763/2020.2.1469]
27. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Sci Eng Ethics. 2020;26(6):3333-3361. [FREE Full text] [doi: 10.1007/s11948-020-00276-4] [Medline: 33196975]
28. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. European Commission. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [accessed 2024-01-16]
29. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI Soc. 2023;38(2):549-563. [FREE Full text] [doi: 10.1007/s00146-022-01455-6] [Medline: 35615443]
30. The evolution of generative AI: a deep dive into the life cycle and training of advanced language models? LinkedIn. URL: https://www.linkedin.com/pulse/evolution-generative-ai-deep-dive-life-cycle-training-aritra-ghosh/ [accessed 2023-09-20]

31. Sachdeva PS, Barreto R, von VC, Kennedy CJ. Assessing annotator identity sensitivity via item response theory: a case study in a hate speech corpus. USA. Association for Computing Machinery; 2022. Presented at: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; 2022 June 21-24; Seoul Republic of Korea. [doi: 10.1145/3531146.3533216]
32. Chan A. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. AI Ethics. 2023;3(1):53-64. [doi: 10.1007/s43681-022-00148-6]
33. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? USA. Association for Computing Machinery; 2021. Presented at: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 March 3-10:610-623; Virtual Event, Canada. [doi: 10.1145/3442188.3445922]
34. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Min Knowl Discov. 2020;10(3):e1356. [doi: 10.1002/widm.1356]
35. Sun T, Gaut A, Tang S, Huang Y, ElSherief M, Zhao J, et al. Mitigating gender bias in natural language processing: literature review. Association for Computational Linguistics; 2019. Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 July 28- August 2:1630-1640; Florence, Italy. [doi: 10.18653/v1/p19-1159]
36. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Lang Linguist Compass. 2021;15(8):e12432. [FREE Full text] [doi: 10.1111/lnc3.12432] [Medline: 35864931]
37. Steed R, Caliskan A. Image representations learned with unsupervised pre-training contain human-like biases. 2021. Presented at: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 March 3-10:701-713; Virtual Event, Canada. [doi: 10.1145/3442188.3445932]
38. How it feels to be sexually objectified by an AI. MIT Technol Rev. URL: https://www.technologyreview.com/2022/12/13/1064810/how-it-feels-to-be-sexually-objectified-by-an-ai/ [accessed 2023-09-20]
39. Castaneda J, Jover A, Calvet L, Yanes S, Juan A, Sainz M. Dealing with gender bias issues in data-algorithmic processes: a social-statistical perspective. Algorithms Multidisciplinary Digital Publishing Institute. 2022;15(9):303. [doi: 10.3390/a15090303]
40. Wellner G, Rothman T. Feminist AI: can we expect our AI systems to become feminist? Philos Technol. 2020;33(2):191-205. [doi: 10.1007/s13347-019-00352-z]
41. Zhou J, Zhang Y, Luo Q, Parker A, De CM. Synthetic lies: understanding AI-generated misinformation and evaluating algorithmic and human solutions. ACM; 2023. Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023 April 23 - 28:1-20; Hamburg, Germany. [doi: 10.1145/3544548.3581318]
42. Vinay R, Spitale G, Biller-Andorno N, Germani F. Emotional manipulation through prompt engineering amplifies disinformation generation in AI large language models. Computer Science > Artificial Intelligence. 2024:1-14. [doi: 10.48550/arXiv.2403.03550]
43. Four years later, AI language dataset created by brown graduate students goes viral. Brown Univ. 2023. URL: https://www.brown.edu/news/2023-04-25/open-web-text [accessed 2023-09-20]
44. Patenaude J, Legault G, Beauvais J, Bernier L, Béland J, Boissy P, et al. Framework for the Analysis of Nanotechnologies? Impacts and ethical acceptability: basis of an interdisciplinary approach to assessing novel technologies. Sci Eng Ethics. 2025;21(2):293-315. [doi: 10.1007/s11948-014-9543-y]
45. Taebi B. Bridging the gap between social acceptance and ethical acceptability. Risk Anal. 2017;37(10):1817-1827. [doi: 10.1111/risa.12734] [Medline: 27862106]
46. Hacker P. A legal framework for AI training data from first principles to the artificial intelligence act. Law Innov Technol Routledge. 2021;13(2):257-301. [doi: 10.1080/17579961.2021.1977219]
47. Artificial intelligence act: deal on comprehensive rules for trustworthy AI. News | European Parliament. 2023. URL: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai [accessed 2024-01-17]
48. Goldstein J, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K. Generative language models and automated influence operations: emerging threats and potential mitigations. arXiv. 2023. URL: http://arxiv.org/abs/2301.04246 [accessed 2023-09-20]
49. Lessons Learned on Language Model Safety and Misuse. URL: https://openai.com/research/language-model-safety-and-misuse [accessed 2023-09-20]
50. Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv. URL: http://arxiv.org/abs/2209.07858 [accessed 2023-09-20]
51. The AI Detection Arms Race Is On | WIRED. URL: https://www.wired.com/story/ai-detection-chat-gpt-college-students/ [accessed 2023-09-20]
52. Smith T. Arms race instability and war. J Confl Resolut SAGE Publications Inc. 1980;24(2):253-284. [doi: 10.1177/002200278002400204]
53. ChatGPT and the fight against disinformation: how AI is changing the game. Culturico. URL: https://culturico.com/2023/03/04/chatgpt-and-the-fight-against-disinformation-how-ai-is-changing-the-game/ [accessed 2023-09-20]

54. Germani F, Biller-Andorno N. How to counter the anti-vaccine rhetoric: filling information voids and building resilience. Hum Vaccin Immunother. 2022;18(6):2095825. [FREE Full text] [doi: 10.1080/21645515.2022.2095825] [Medline: 35802046]
55. Convention on cybercrime ETS - No. 185. 2001. Council of Europe. 2001. URL: https://rm.coe.int/1680081561 [accessed 2024-09-17]
56. How Digital Identity can Protect Against Misuse of AI. URL: https://oneid.uk/news-and-events/how-digital-identity-can-protect-against-misuse-of-ai [accessed 2023-09-20]
57. Ahmad W, Berg R, Kim S. Combating Fake News with Digital Identity Verification. URL: https://groups.csail.mit.edu/mac/classes/6.805/student-papers/fall17-papers/FakeNews.pdf [accessed 2024-09-17]
58. DeVerna MR, Yan HY, Yang KC, Menczer F. Artificial intelligence is ineffective and potentially harmful for fact checking. arXiv. 2023. URL: http://arxiv.org/abs/2308.10800 [accessed 2023-09-20]
59. Sebastian G. Exploring ethical implications of ChatGPT and other AI Chatbots and regulation of disinformation propagation. SSRN. 2023:1-16. [doi: 10.2139/ssrn.4461801]
60. About Community Notes on X | X Help. URL: https://help.twitter.com/en/using-x/community-notes [accessed 2023-09-21]
61. Directorate general for parliamentary research services. Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism. European Parliament. LU. Publications Office; 2019. URL: https://data.europa.eu/doi/10.2861/003689 [accessed 2023-09-21]
62. Meyer T. Regulating Disinformation with Artificial Intelligence? URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf [accessed 2024-09-17]
63. Harris ED. Governance of Dual-Use Technologies. URL: https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/3 [accessed 2024-09-17]
64. Appedu S, Hensley MK. Problematizing the role of information literacy in disinformation, dialogue, the healing of democracy. In: Sietz B, editor. Inf Lit Time Transform. Michigan. LOEX Press; 2021.
65. Ringing the Alarm Bell with Federico Germani. URL: https://www.andybusam.com/ringing-the-alarm-bell-with-federico-germani/ [accessed 2023-09-21]
66. Redaelli S, Biller-Andorno N, Gloeckler S, Brown J, Spitale G, Germani F. Mastering critical thinking skills is strongly associated with the ability to recognize fakeness and misinformation. SocArXiv (OSF). 2024. [doi: 10.31235/osf.io/hsz6a]
67. Jones-Jang SM, Mortensen T, Liu J. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. Am Behav Sci. 2019;65(2):371-388. [doi: 10.1177/0002764219869406]
68. De PS, Heravi B. Information literacy and fake news: how the field of librarianship can help combat the epidemic of fake news. J Acad Librariansh. 2020;46(5):102218. [doi: 10.1016/j.acalib.2020.102218]
69. Willingham DT. Ask the cognitive scientist: how can educators teach critical thinking? Am Educ American Federation of Teachers, AFL-CIO. 2020;3(41):44.
70. Gaillard S, Oláh ZA, Venmans S, Burke M. Countering the cognitive, linguistic, and psychological underpinnings behind susceptibility to fake news: a review of current literature with special focus on the role of age and digital literacy. Front Commun. 2021. URL: https://www.frontiersin.org/articles/10.3389/fcomm.2021.661801 [accessed 2023-09-26]
71. Allner IB. Teaching of Information Literacy: Collaboration Between Teaching Faculty and Librarians. US. BiblioBazaar; 2011.
72. Johnston B, Webber S. Information literacy in higher education: a review and case study. Stud High Educ Routledge. 2003;28(3):335-352. [doi: 10.1080/03075070309295]
73. Burclaff N, Johnson C. Teaching Information Literacy via Social Media: An Exploration of Connectivism. URL: https://www.researchgate.net/publication/316187027_Teaching_Information_Literacy_via_Social_Media_An_Exploration_of_Connectivism [accessed 2024-09-17]
74. Fake news created by artificial intelligence is difficult to recognize. They seem more credible to Internet users than messages created by humans. Bizness. URL: http://biznes.newseria.pl/news/fake-newsy-stworzone-przez,p919781558 [accessed 2023-09-22]
75. Tiernan P, Costello E, Donlon E, Parysz M, Scriney M. Information and media literacy in the age of AI: options for the future. Educ Sci Multidisciplinary Digital Publishing Institute. 2023;13(9):906. [doi: 10.3390/educsci13090906]
76. Hwang Y, Ryu JY, Jeong S. Effects of disinformation using deepfake: the protective effect of media literacy education. Cyberpsychol Behav Soc Netw. 2021;24(3):188-193. [doi: 10.1089/cyber.2020.0174] [Medline: 33646021]
77. WHO Kicks off Deliberations on Ethical Framework and Tools for Social Listening and Infodemic Management. URL: https://www.who.int/news/item/10-02-2023-who-kicks-off-deliberations-on-ethical-framework-and-tools-for-social-listening-and-infodemic-management [accessed 2023-09-22]
78. Pause giant AI experiments: an open letter. Future Life Inst. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [accessed 2023-10-02]

Abbreviations

AI: artificial intelligence

Edited by K El Emam, B Malin, A Blasimme; submitted 09.10.23; peer-reviewed by E Pertwee, S Gordon, E Wilhelm; comments to author 13.01.24; revised version received 22.01.24; accepted 28.07.24; published 15.10.24

Please cite as:
Germani F, Spitale G, Biller-Andorno N
The Dual Nature of AI in Information Dissemination: Ethical Considerations
JMIR AI 2024;3:e53505
URL: https://ai.jmir.org/2024/1/e53505
doi: 10.2196/53505
PMID: 39405099