Nursing Inquiry

WILEY

Nursing *Inquiry*

ORIGINAL ARTICLE OPEN ACCESS

Is Generative AI Increasing the Risk for Technology-Mediated Trauma Among Vulnerable Populations?

Abdul-Fatawu Abdulai

School of Nursing, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence: Abdul-Fatawu Abdulai (fatawu.abdulai@ubc.ca)

Received: 7 June 2024 | **Revised:** 28 October 2024 | **Accepted:** 31 October 2024

Funding: The author received no specific funding for this work.

Keywords: artificial intelligence | digital health | emotional trauma | generative AI | technology-mediated trauma

ABSTRACT

The proliferation of Generative Artificial Intelligence (Generative AI) has led to an increased reliance on AI-generated content for designing and deploying digital health interventions. While generative AI has the potential to facilitate and automate healthcare, there are concerns that AI-generated content and AI-generated health advice could trigger, perpetuate, or exacerbate prior traumatic experiences among vulnerable populations. In this discussion article, I examined how generative-AI-powered digital health interventions could trigger, perpetuate, or exacerbate emotional trauma among vulnerable populations who rely on digital health interventions as complementary or alternative sources of seeking health services or information. I then proposed actionable strategies for mitigating AI-generated trauma in the context of digital health interventions. The arguments raised in this article are expected to shift the focus of AI practitioners against prioritizing dominant narratives in AI algorithms into seriously considering the needs of vulnerable minority groups who are at the greatest risk for trauma but are often invisible in AI data sets, AI algorithms, and their resultant technologies.

1 | Introduction

1.1 | What Is Generative AI

Generative AI, encompassing models like ChatGPT, Google's Bard, and others, is an emerging sub-field of AI that uses machine learning models to generate high-quality text, images, and other content based on the data they are trained on (IBM Research 2021). While earlier AI models were limited to analyzing, interpreting, and predicting scenarios based on existing data, generative AI can create new content (Houssami et al. 2019). Generative AI automatically learns patterns and structures from texts, images, sounds, animations, models, or other media inputs to produce new ones with similar characteristics. The introduction of generative AI has brought accelerated advancements and is gaining prominence in various fields including healthcare (Raza, Venkatesh, and Kvedar 2024; Reddy 2024). The ability to generate new content, plus the accessibility and ease of use of generative AI, has led to a surge in use by both patients and providers (Carrie 2024).

1.2 | Generative AI in Healthcare

The proliferation of Generative AI has led to an increased reliance on AI-assisted information seeking, clinical decision-making, diagnosis, routine monitoring, treatment recommendations, and post-treatment monitoring among healthcare providers and patients (Carrie 2024). Generative AI is also used in the design, development, and deployment of digital health interventions (Raza, Venkatesh, and Kvedar 2024). According to the WHO, generative AI can be used for creating content and images for digital health user interfaces, improving the quality of existing ones, or enhancing the resolution of images in digital

health platforms (WHO 2024). Generative AI has the potential to streamline the development of digital health interventions by facilitating the coding process and improving the development of computational solutions for patients. The potential of generative AI in digital health was demonstrated in a study that used ChatGPT to develop a digital health intervention for supporting patient engagement and adherence to digital diabetes prevention programs (Rodriguez et al. 2024). In this project, ChatGPT was not only used in developing the codes for the intervention but also supported the conceptualization of the product, content development, software requirement generation, and software design (Rodriguez et al. 2024). Beyond assisting with the design and development of digital health, generative AI can also help patients become more autonomous by facilitating individualized assessment of data and symptoms at a faster pace than humans can do (Meskó and Topol 2023). For instance, chatbots can provide patients with 24-h access to health information including symptom assessment, medication reminders, counseling information, and appointment scheduling (Clark and Bailey 2024).

While generative AI has great promise in the development of digital health and delivery of healthcare, there are concerns that advancements in this technology could increase the risk of technology-mediated trauma among technology users, particularly vulnerable populations. In this article, vulnerable populations are used to refer to people and groups including those with intellectual disabilities, sex workers, 2SLGBTQI+ people, racialized people, indigenous, youth as well as elderly patients, and religious groups who are most at risk of technology-facilitated abuse. This definition is adapted from the government of Canada's 2018 definition of vulnerable population as encompassing marginalized identities including immigrants, children, Indigenous, newcomers, members of LGBTQ2+ communities, language minority communities, people living with mental illness, people living with homelessness, persons with disabilities, etc. (Government of Canada 2018). While all technology users, including the privileged, might be faced with technology-facilitated abuse, it is important to state that vulnerable populations may be more impacted (Chen et al. 2022; Venkatasubramanian and Ranalli 2022). The risk of technology-mediated trauma among vulnerable populations is heightened because of prior traumatic experiences at the personal, structural, and systemic levels in the healthcare systems (Haimson et al. 2020; Matthews et al. 2017; Simko et al. 2018).

Technology-based interventions can trigger or exacerbate such prior traumatic experiences if used inappropriately. For instance, various technology-based interventions such as mobile applications, spyware tools, and the Internet of Things could be used to cause harm that may result in traumatic symptoms including anxiety, panic attacks, fear, anger, and irritability (MacLure and Jones 2021). Other technology-facilitated cycles of abuse can take the form of online sexual harassment, cyberstalking, image-based exploitation, and the use of manipulative technologies services to coerce people into unwanted sexual acts (Henry and Powell 2018). Given that vulnerable populations may have a disproportionate rate of trauma (Newland et al. 2022), any negative effects of digital technologies could trigger or exacerbate any prior traumatic experiences. This emerging threat of technology-mediated trauma is particularly concerning for people who rely on AI-powered digital health interventions as alternative or complementary sources of seeking health information or services. To contextualize the discussions and understand how generative AI can trigger, perpetuate, and/or exacerbate trauma, I first provided an overview of trauma and how trauma occurs in the context of digital health technologies.

1.3 | Trauma

In this article, trauma is used to describe the challenging emotional and psychological consequences associated with an individual after going through a distressful event (Dupont 1998). This definition of trauma is specific to emotional and psychological trauma, which is different from physical trauma (i.e., injuries or body wounds produced by sudden physical injury from impact, violence, or accident). Available evidence suggests that approximately 70% of people across the globe will experience some form of trauma during their lifetime (Knipscheer et al. 2020). This number is estimated to be high among vulnerable populations who are already faced with systematic injustices and inequitable access to healthcare services (Tebes et al. 2019). Trauma resulting from the use of digital health technologies (herein referred to as technology-mediated trauma) is an emerging threat that has a profound impact on the lives of technology users, particularly among vulnerable groups (Chen et al. 2022). Given these high rates of trauma and the concomitant increase in the use of digital health technologies, it means many technology users might be survivors of trauma (Wong et al. 2021). For vulnerable populations who might have experienced prior traumatic events, any additional negative effects of health technologies might trigger, perpetuate, or exacerbate prior traumatic experiences – further exacerbating health inequities for those with the greatest need.

1.4 | Technology-Mediated Trauma in Digital Health

While there is not yet a universally accepted definition of technology-mediated trauma, it is generally conceptualized as technology-enabled activities, interface appearances, or features that leave a person in emotional distress following the use of digital technology (Chen et al. 2022). Technology-mediated trauma may occur inadvertently through interface features that remind users of prior traumatic circumstances (Holloway et al. 2014; Musetti et al. 2021). For instance, technologies that expose people to online scams, sextortion, identity theft, and cyberbullying could result in short-term or long-term emotional trauma, and sometimes, suicidal thoughts for people (Kraft 2006; Schüz and Urban 2020). For survivors of trauma like victims of childhood sexual abuse, viewing certain sexually explicit content that they consider to be inappropriate can re-traumatize them and further create technology-facilitated cycles of trauma or re-traumatization each time the content is viewed (Regehr, Birze, and Regehr 2022). Even seemingly neutral user interface design elements developed in line with established design guidelines (e.g., the use of different colors and images) can trigger emotional trauma in people with negative prior

experiences or a negative attitude toward the said elements (Shaughnessy et al. 2022). Bonell et al. (2015) termed these sorts of occurrences as the "dark logic" of public health interventions while Ziebland, Hyde, and Powell (2021) also described it as the "paradoxical consequences of technologies" where interventions designed for treating a particular condition inadvertently deliver a contrary outcome.

Aside from the inadvertent effects of technology on trauma, trauma can also be orchestrated intentionally through negative online activities. For instance, Digital technologies can trigger emotional trauma through interface features that can be used to harass, coerce, or stalk someone (Freed et al. 2018). These sorts of abuses can range from intimate image abuse, and the use of geolocation features for stalking, to the use of Internet of Things devices to manipulate people into doubting their sanity (Regehr, Birze, and Regehr 2022). Perpetrators of intimate partner abuse and domestic violence can install or place Global Positioning Systems (GPS) devices to track their victims and possibly cause harm (Straw and Tanczer 2023). Cyberbullying perpetrated via online social networking sites such as negative forum posts and comments has been shown to result in emotional trauma among patients with prior histories of trauma (Habib et al. 2021; Hong et al. 2020). Other forms of communication channels on digital platforms including chatrooms, user subscription channels, and forums/bulletin boards may be used unscrupulously to victimize and expose users in a manner that could result in emotional trauma (Henry, Flynn, and Powell 2020; Machimbarrena et al. 2018).

The preceding evidence suggests that technology-mediated trauma is not a new phenomenon but a set of events that are associated with technology use. With the innovative and widespread use of generative AI in social media (Sheikh and Rogers 2024), I argue that technology-mediated abuse may not only be on the increase but could also be perpetrated by lay and less sophisticated people with ill motives. For instance, generative AI can exacerbate trauma by introducing new, highly personalized, and invasive forms of harassment (UK Council for Internet Safety 2019). Also, biased AI models can enhance technology-facilitated abuse by enabling the creation of highly realistic and deceptive digital content, such as deepfake videos, images, and audio recordings, which can be used to manipulate, impersonate, or defame victims (Stonard et al. 2017). This advanced capability can significantly intensify the victim's sense of helplessness and violation, as the fabricated content can be indistinguishable from reality, making it harder to disprove and stop its spread. The sophistication and reach of generative AI can amplify the psychological impact, leading to severe anxiety, fear, and trust issues (Lucas 2022). The use of AI-enabled deepfake videos has led to unprecedented growth in image-based intimate sexual abuse with profound emotional and social impacts on the victims' lives (Lucas 2022). A recent study shows that deepfake videos can be used to enact sexual violence and such acts could be particularly harmful and could result in endless cycles of abuse for the victims (Rousay Victoria 2023). The enduring and replicable nature of AI-generated content means that the abuse can persist indefinitely, further entrenching the trauma and its long-term psychological effects, such as PTSD, depression, and a diminished sense of personal security (Regehr, Birze, and Regehr 2022; Sheikh and Rogers 2024). For vulnerable populations who already experience trauma at the personal, structural, and systemic levels in the healthcare systems (Haimson et al. 2020; Matthews et al. 2017; Simko et al. 2018), such abuses can cause long-lasting or permanent emotional trauma that could reoccur anytime such videos or images are brought to their attention. In such situations, victims may experience an overwhelming sense of loss of control over their digital and real-world identities (Rousay Victoria 2023). With the increasing use of AI-powered technologies among vulnerable populations including survivors of trauma (Venkatasubramanian and Ranalli 2022), any potential negative consequence of such technologies could exacerbate trauma and widen health inequities among vulnerable populations who have the most needs but face the greatest barriers in access to care. In the rest of this article, I explained how AI algorithms and their resultant technologies can trigger or perpetuate harmful consequences including trauma among historically vulnerable populations. I then proposed actionable strategies for ensuring safer and trauma-informed AI-powered digital health technology interventions.

2 | How Generative AI Increases the Risk for Technology-Mediated Trauma

Generative AI is seen as a promising tool that can speed up the design and use of digital health interventions and facilitate autonomous decision-making for patients (Rodriguez et al. 2024). The opportunities presented by generative AI for automating healthcare are both exciting and problematic at the same time. Despite its vast potential, there are concerns that the use of generative AI can increase the risk for technology-mediated trauma – posing a risk to people with prior traumatic experiences who rely on digital health for information and services. Traumatic experiences via generative AI could occur in several ways including the use of underrepresented and biased data sets used in training AI algorithms, the provision of inaccurate health information, and the intentional creation of malicious AI models for spreading disinformation.

2.1 | Underrepresented, Biased, and Inaccurate Data Sets Used in Training AI algorithms

First of all, it is important to understand that current AI algorithms rely largely on data from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich, Heine, and Norenzayan 2010). Vulnerable minority groups who are susceptible to trauma are often underrepresented in data sets that are used in training AI models (Shanklin et al. 2022). With such underrepresentation, AI algorithms may often not recognize minority and marginalized groups whose concerns are often perceived to occur outside the dominant narratives. For instance, one study found that a widely used patient prioritization AI algorithm was more likely to prioritize white patients over black patients with the same level of health needs (Obermeyer et al. 2019). Another study on the deployment of AI-enabled medical imaging was found to exacerbate existing biases among Hispanic female patients and could potentially lead to unequal access to medical care for underserved

populations like women (Seyyed-Kalantari et al. 2021). Such biases may not only perpetuate existing health inequities but also downplay or exacerbate concerns that could trigger or perpetuate trauma among marginalized populations (Larasati 2023).

The lack of recognition of different marginalized identities in AI algorithms and AI-powered digital technologies may reinforce a feeling of lack of belongingness and create harmful consequences including anxiety, anger, lack of control, and self-inadequacy among marginalized populations who use AI-powered technology interventions (Holmes 2018). Also, biased online information on marginalized populations could lead to problematic representation of minority groups while amplifying harmful social biases and negative stereotypes in generative AI-powered technologies (Hastings 2024). Also, AI algorithms that are trained on data sets that contain inadequate data on gender, racial, or ethnic representation may exhibit the same racist and gender biases and reinforce harmful stereotypes as their source (Obermeyer et al. 2019; Lamensch 2024). For instance, a predictive generative AI algorithm that predominantly contains images of white skin was found to produce biased and inaccurate images of dark-skinned people (Adamson and Smith 2018). Technologies resulting from such algorithms can either exacerbate prior trauma or introduce new forms of trauma by preventing people from seeking help with the intervention (Eggleston and Noel 2024). Digital health interventions that use biased AI-generated content and images may also appear as the exclusion of people who experience marginalization and a lack of attention to the safety and care of these populations (Holmes 2018).

In addition to the biases in AI algorithms, generative AI may also use data sets that may contain inappropriate, or harmful information about certain minority groups. While some AI companies try to validate their data sources and remove inappropriate data sources (Biessmann et al. 2021), those data sources might still contain information that could be harmful or at best inappropriate for specific user groups. Therefore, AI-powered technologies resulting from inaccurate data sources might result in user interface content that could be traumatizing for vulnerable minority groups who find such content as either undesirable, reinforcing negative stereotypes, or reminding them of prior traumatic events (Regehr, Birze, and Regehr 2022). For instance, AI algorithms that are trained on internet-based information may use "naked images" of people to create or replicate hypersexualized images that could be harmful to sexual minority groups. These harmful misrepresentations are more common in sexual health-related technologies where hypersexualized data from the internet are used to create and propagate negative images of minority identities including 2SLGBTQI, blacks, and religious minority groups (Lamensch 2024). For instance, a recent study found ChatGPT-4 to depict problematic representation and harmful social biases that carry considerable risks of triggering, exacerbating, and perpetuating harmful consequences among minority groups including blacks, 2SLGBTQI people, indigenous, racial, and religious minorities (Zack et al. 2024). These AI-mediated cultural misrepresentations are especially concerning when AI algorithms cannot distinguish between appropriate cultural appropriation and negative cultural stereotyping. Cultural appropriation is the representation of cultural practices or experiences by "outsiders," while negative cultural stereotyping refers to widely held beliefs (mostly in a negative light) about a particular social group (Kashima 2008; Matthes 2016). In addition to the biases and inaccurate representation of minority groups in AI algorithms, there is also a lack of diversity within the teams that design and deploy AI algorithms (Green, Murphy, and Robinson 2024). Research from the AI Now Institute (https://ainowinstitute.org/) highlighted that the majority of AI developers and researchers were white cis-gender males (AI Now Institute 2018). The powerful and often non-diverse technology industry tends to hold a narrow view of marginalized populations or may not even understand trauma. As a result, such teams may not understand the implications of design decisions on marginalized populations and how well-intentioned design decisions could inadvertently re-traumatize minority identities.

2.2 | Inappropriate Health Advice via AI-Powered Chatbots

One other area in which generative AI can foment trauma is through AI-powered technologies such as chatbots used in delivering healthcare (Boucher et al. 2021). There has been a recent increase in the use of chatbots to counsel people, including survivors of abuse, and also to connect people experiencing high rates of trauma and psychological distress with healthcare providers (Ngũnjiri et al. 2023). While such chatbots might be easily accessible and convenient to use, the inaccuracies and underrepresentations in AI data sets and AI algorithms might lead to situations where AI-powered chatbots use words, and statements or give health advice that reminds people of prior traumatic events, thereby re-traumatizing rather than helping them (Coghlan et al. 2023; McMahon and McMahon 2024).

Empathy, which is considered the ability to recognize people's thoughts and feelings and to respond with appropriate emotions, is an essential element for survivors of trauma (Greenberg et al. 2018). For survivors of trauma, empathy and compassionate engagement are necessary for fostering connection, building resilience, empowerment, and gaining coping strategies after going through a distressful event (Nugent, Sumner, and Amstadter 2014). While recent studies have demonstrated the ability of ChatGPT to recognize human emotions and produce empathetic responses to queries (Elyoseph et al. 2023), generative AI does not yet fully possess the human capabilities for conveying patient-specific emotional demands and other traits that require critical thinking and emotional intelligence. Even when AI models can communicate with empathy, the algorithms may not understand the patients' context or may not contain background information to be able to tailor responses or distinguish between what could be helpful or traumatizing to each patient's circumstances. For instance, ChatGPT may not understand the traumatic past of a patient who uses an AI-powered chatbot to seek information on abortion. In healthcare and nursing, in particular, each individual is unique, and generative AI in its current state may not be able to respond to the uniqueness of each patient (Katims 1995). Other crucial human communication traits such as active listening, eye contact, and

portraying a sense of connectedness with patients are important for trauma survivors but not inherent in current generative AI models. Without the ability to convey such traits, AI-powered technologies could end up being distressful to vulnerable populations that urgently need emotional connection and empathetic communication to be able to deal with their traumatic past.

2.3 | Intentional Use of Malicious AI Models

Another way that generative AI increases the risks of technology-mediated trauma is through the intentional development of malicious AI algorithms that can be used to target vulnerable populations including survivors of trauma. Despite the proactive steps by companies like Open AI and Google Bard to mitigate the nefarious use of generative AI, malicious AI models have been developed and will continue to be created (Ferrara 2024). These adversarial AI models highlight the potential threats generative AI poses to the health and well-being of vulnerable populations. Malicious AI models such as WormGPT and PoisonGPT, for instance, were created as "black hat" alternatives to standard ChatGPT (McGowan 2023). These AI models can be used to produce deceptive narratives, fake images, and deepfake videos that could trigger and perpetuate prior traumatic events with harmful consequences (Maras and Alexandrou 2019). Deepfakes are videos that have been convincingly manipulated to represent someone and spread misinformation (Maras and Alexandrou 2019). For instance, WormGPT can be leveraged to create malicious content, images, or information targeting vulnerable populations – triggering and perpetuating prior traumatic experiences. Another concerning scenario is the potential modification of existing AI algorithms to spread misinformation and disinformation through open-source coding. For instance, generative models such as PoisonGPT can be used to generate images, videos, and audio content to vividly and persuasively convey misinformation about a particular group or individual (Xu, Fan, and Kankanhalli 2023). This threat, in particular, is real because of the use of unregulated and unaudited internet information/data that are used in training these adversarial AI models (Martínez et al. 2024). While AI-generated misinformation can be used to target people with any disease condition, they are more likely to be targeted at people with sensitive and potentially embarrassing health conditions or people seeking politically sensitive health services (e.g., abortion) (McMahon and McMahon 2024; Xu, Fan, and Kankanhalli 2023). For instance, ChatGPT was found to provide abortion-related information that overstated the risk associated with self-managed abortion in a manner that contradicts current evidence, perpetuates misinformation, and amplifies the stigma associated with abortion (McMahon and McMahon 2024).

Given the increasing adoption of generative AI in designing digital health technologies and automating patient decision-making, the issues raised in this article should be concerning to AI researchers and AI practitioners. It is also important to note that generative AI in healthcare is still in its infancy and, therefore, its effects on trauma might not have been widely noticed or reported yet. Nevertheless, AI researchers and AI practitioners must be aware of these concerns and take actionable strategies to prevent traumatic consequences in future AI algorithms and their resultant technologies.

3 | Potential Actionable Strategies Toward Trauma-Informed AI Technologies

Given the increasing threats of generative AI, researchers and AI practitioners are exploring various ways of designing safer, trustworthy, and responsible AI. For instance, the European Commission developed ethical guidelines for developing responsible and trustworthy AI (Lemonne 2018). Also, tech companies such as Google and agencies like the Responsible AI Institute developed guidelines for creating, procuring, and deploying safe and trustworthy AI systems (Responsible AI Institute 2020). These guidelines are generally meant to promote technical robustness, safety, transparency, and diversity, ensure nondiscrimination, and enhance fairness and accountability in AI technologies. While these guidelines might be gravitating toward a trustworthy and safer AI, it must be noted that they are largely high-level recommendations, not specific to the needs of marginalized populations, and are therefore inadequate in mitigating technology-mediated trauma in AI-powered digital health interventions.

To design AI-powered digital health technologies with the least chance of triggering or perpetuating trauma, health-related AI researchers and industry partners should undergo trauma-informed training to have a better understanding of trauma, the possible trigger factors, and its effects on technology users, particularly vulnerable populations. While trauma-informed care practices are currently being integrated into agile software development and virtual reality programs (Anton Skornyakov 2023; Matt 2024), it is not evident if software engineers in the technology industry, particularly those who design health-related AI models are undergoing any training opportunities before developing AI algorithms, particularly for populations who face existing vulnerabilities in society. To make AI-powered technologies less traumatizing to the end user, I propose that industry partners and AI practitioners must engage with trauma experts and possibly survivors of trauma during the design and development phases to ensure a deep understanding of the needs of trauma survivors. Such engagement would not only help to audit, identify, and address data sources that can foment trauma but also ensure a sense of ownership of AI technology by end-user patients.

While patient engagement is crucial in mitigating technology-mediated trauma, it is important to state that some engagement approaches could re-traumatize some vulnerable groups (Knipscheer et al. 2020). Also, some vulnerable patients may not be in the position or have the capacity to communicate their needs or indicate what could be traumatizing to them. In such instances, health professionals like nurses could advocate for the needs of the patient population to be addressed in AI algorithms and ensure that vulnerable patients are not re-traumatized in an attempt to help them (Nsiah, Siakwa, and Ninnoni 2019). In addition to patient engagement, AI practitioners also need some standardized guidelines to stay abreast with information on their patient population and current issues that could re-traumatize them. This could be like an extension of existing design guidelines to allow for the development of technologies to meet the needs of marginalized populations (Avellan, Sharma, and Turunen 2020; Tengkawan, Agnihotri, and Minhas 2022). Furthermore, inclusive design guidelines

like integrating exit buttons, having simple fonts, reducing cognitive load, limiting the number of navigational options, and providing options for support could be applied in developing AI-powered technologies to limit the potential for distress caused by difficult-to-use interfaces (Kelly, Lauren, and Nguyen 2021; Venkatasubramanian and Ranalli 2022). Indeed, domain-specific design guidelines have been developed for other equally difficult topics like stigma (Abdulai et al. 2021, 2023). Therefore, it is only proper that technology design guidelines tailored to trauma for vulnerable populations are also developed. Such guidelines would not only enhance the user experience for trauma survivors but also promote inclusivity and accessibility – making technology interventions more effective and empathetic for a diverse range of users.

Another concern that needs to be addressed to prevent technology-facilitated abuse is diversity. It has been suggested that AI researchers and developers are mostly white men (AI Now Institute 2018). I contend that it could be challenging to increase minority groups' representation in AI research and development teams in the short term. In the meantime, AI developers should recognize the diversity of potential users, appreciate their lived experiences, and develop AI models to accurately reflect differences in sex, gender, race, education, religion, skills, and disability (Ulnicane 2024). Such appreciation might help AI developers develop models that adapt and respond to the queries of different patient populations without introducing biases or reinforcing stereotypes, thereby addressing the generality of AI-generated responses. For instance, AI researchers at Stanford University developed an AI algorithm that was able to discern people's sexual orientation on the dating site with remarkable accuracy (Wang and Kosinski 2018). These sorts of algorithms could prevent technology-mediated trauma often caused by misrepresentation or non-representation of sexual minority groups (Eggleston and Noel 2024). Recognizing and integrating diversity is important because non-diverse AI algorithms could result in technologies that essentially reject nonnormative practices and rather project beliefs that could be oppressive, traumatizing, and negatively impact the safety of minority groups who face other vulnerabilities in life (Nicki 2001).

Research shows that AI-powered technologies like Chatbots could result in technology-mediated trauma if the advice offered reinforces prior traumatic experiences (Bendig et al. 2019). As chatbots are being integrated with electronic health records (Clark and Bailey 2024), urgent steps are needed to evaluate the clinical utility of these technologies and how they can adapt to the specific circumstances of patients, particularly vulnerable groups who are prone to trauma. For example, EPIC and Cerner (Electronic Health Records companies) partnered with Microsoft Open AI to integrate chatbots into their Electronic Health Records – enabling EHRs to auto-draft responses to patient queries, order medications, and schedule follow-up appointments (Microsoft News Center 2023). Given the potential risk that the misuse of these technologies poses to vulnerable groups, appropriate actions must be taken to evaluate their clinical value for minority groups before being implemented for patient care. Such evaluation could adopt the framework developed by Wornow et al. (2023) for evaluating AI models in healthcare settings. The framework comprises six criteria including predictive performance, data labeling, model development, emergent clinical applications, multimodality, and novel human–AI user interfaces.

Beyond establishing clinical utility for minority groups, I propose that AI algorithms should also contain context-specific information and human values. This is consistent with prior and recent calls for integrating human and ethical values in developing technology interventions (Barnard and Sandelowski 2001; Duarte and Baranauskas 2016). It has already been established that ChatGPT can recognize human emotions and produce empathetic responses to queries (Elyoseph et al. 2023). Digital empathy can be enhanced by having tech developers work closely with nurses in developing AI-powered technologies including chatbots. Nurses are particularly skilled in empathy and individualized care and could help in developing AI-powered chatbots that not only communicate with empathy but also provide individualized and context-specific responses (Rey Velasco et al. 2024). Also, given the inherent variability in what people perceive as correct responses and the fact that medical AI algorithms are often unclear about the target population, settings, and the handling of missed data (Collins et al. 2024; Wu et al. 2021; Wynants et al. 2020), further efforts should be made to contextualize and provide individual-specific and human-centric responses. Recent developments in AI research show that this is possible as there are growing body of work to make AI more human and individual-centered (Dhamala et al. 2021; Nadeem, Bethke, and Reddy 2021; Ouyang et al. 2024). These efforts should be primarily aimed at identifying, integrating, and aligning human values in AI models. Furthermore, AI-powered chatbots should provide real-time responses by relying on the input and context of the patient (Meskó and Topol 2023). Such AI-enabled adaptive systems have been successfully used in education and learning systems and could prove useful for health-related technologies (Kabudi, Pappas, and Olsen 2021).

In addition to the lower-level actions toward mitigating technology-mediated trauma, I propose that governments, the private sector, industry leaders, and academic researchers on AI should partner to discuss governance mechanisms that mitigate the risks and potential harm of AI while harnessing its potential for vulnerable groups. Such governance mechanisms could include collaborative efforts at building transparency, fairness, and accountability in the design, deployment, and clinical validation of AI algorithms (Xu, Bradford, and Garg 2023). There has been a strong and consistent call for transparent and ethical AI to promote equitable use of technology but such practices are not necessarily enforced (Basu, Faghmous, and Doupe 2020; Larsson and Heintz 2020). For instance, while some generative AI models such as Adobe Firely (i.e., a generative AI for creating images) have seen adhering to transparency guidelines by publishing information on the data/images used in training their model, others such as Dall-E, have been accused of lacking transparency in their training data (Marr 2024). This lack of transparency makes it difficult to assess the risks of AI models used in generating images for AI-powered digital health technologies (Marr 2024). Publishing information on the AI development process, including the data sources, model choices, and the algorithms used in designing AI technologies would provide enough information to inform the public about the safety and

potential risks of AI algorithms (Fehr et al. 2024). For already existing AI-powered technologies like chatbots, developers should publish regular transparency reports that highlight the systems' performance, accuracy rate, biases, and error rates as well as disclose the data of population sub-groups that are underrepresented in AI algorithms (Xu, Bradford, and Garg 2023). There is a greater chance that generative AI would improve people's health and well-being if people trust the AI technologies to behave safely, securely, and understandably.

4 | Conclusion

Technology-mediated trauma is an emerging threat that requires urgent attention when developing AI-powered digital health technologies. As generative AI is gradually adopted in healthcare, it is important to assess the potential effects of the technology on trauma and take actionable strategies to mitigate technology-mediated trauma among vulnerable populations including survivors of trauma. In this discussion article, I examined various ways in which AI algorithms and their resultant digital health technologies could result in traumatic consequences and proposed actionable strategies for mitigating technology-mediated trauma among vulnerable populations with prior experiences of trauma. Addressing technology-mediated trauma is important because if we ignore the issue of trauma in designing and deploying AI algorithms, we risk re-traumatizing vulnerable populations who will use such interventions, or at best being unhelpful to them. Addressing the needs of vulnerable populations is crucial because current AI algorithms almost always focus on addressing the needs of 80% of the population while neglecting what they consider "edge cases" or situations that are considered to occur outside the dominant narrative (Hastings 2024). I expect that the arguments raised in this article will sensitize AI practitioners against prioritizing dominant narratives in AI algorithms to prioritizing minority groups, or the so-called "edge cases" in designing and deploying AI algorithms. After all, a design that addresses the needs of vulnerable minority groups would inevitably address the needs of dominant groups. As patients increasingly use AI-powered technologies, we also have a moral obligation to ensure that such interventions are not only mitigating the unintended effects of trauma on users but also empowering users to deal with traumatic situations.

Ethics Statement

This study does not require ethics approval.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

No new data were generated for this manuscript.

References

Abdulai, A.-F., A. F. Howard, H. Noga, P. J. Yong, and L. M. Currie. 2021. "Application of Anti-Stigma Design Heuristics for Usability Inspection." *Studies in Health Technology and Informatics* 284: 239–243. https://doi.org/10.3233/SHTI210715.

Abdulai, A.-F., A. F. Howard, P. J. Yong, and L. M. Currie. 2023. "Defining Destigmatizing Design Guidelines for Use in Sexual Health-Related Digital Technologies: A Delphi Study." *PLOS Digital Health* 2, no. 7: e0000223. https://doi.org/10.1371/journal.pdig.0000223.

Adamson, A. S., and A. Smith. 2018. "Machine Learning and Health Care Disparities in Dermatology." *JAMA Dermatology* 154, no. 11: 1247–1248. https://doi.org/10.1001/jamadermatol.2018.2348.

AI Now Institute. 2018. "AI Now 2018 Report." https://ainowinstitute.org/publication/ai-now-2018-report-2.

Avellan, T., S. Sharma, and M. Turunen (2020). "AI for All: Defining the What, Why, and How of Inclusive AI." Proceedings of the 23rd International Conference on Academic Mindtrek, 142–144. https://doi.org/10.1145/3377290.3377317.

Barnard, A., and M. Sandelowski. 2001. "Technology and Humane Nursing Care: (Ir)Reconcilable or Invented Difference?" *Journal of Advanced Nursing* 34, no. 3: 367–375. https://doi.org/10.1046/j.1365-2648.2001.01768.x.

Basu, S., J. H. Faghmous, and P. Doupe. 2020. "Machine Learning Methods for Precision Medicine Research Designed to Reduce Health Disparities: A Structured Tutorial." *Ethnicity & Disease* 30, no. Suppl 1: 217–228. https://doi.org/10.18865/ed.30.S1.217.

Bendig, E., B. Erb, L. Schulze-Thuesing, and H. Baumeister. 2019. "The Next Generation: Chatbots in Clinical Psychology and Psychotherapy to Foster Mental Health – A Scoping Review." *Verhaltenstherapie* 32, no. Suppl. 1: 64–76. https://doi.org/10.1159/000501812.

Biessmann, F., J. Golebiowski, T. Rukat, D. Lange, and P. Schmidt. 2021. "Automated Data Validation in Machine Learning Systems." In *Amazon Science*. https://www.amazon.science/publications/automated-data-validation-in-machine-learning-systems.

Bonell, C., F. Jamal, G. J. Melendez-Torres, and S. Cummins. 2015. "'Dark Logic': Theorising the Harmful Consequences of Public Health Interventions." *Journal of Epidemiology and Community Health* 69, no. 1: 95–98. https://doi.org/10.1136/jech-2014-204671.

Boucher, E. M., N. R. Harake, H. E. Ward, et al. 2021. "Artificially Intelligent Chatbots in Digital Mental Health Interventions: A Review." *Expert Review of Medical Devices* 18, no. sup1: 37–49. https://doi.org/10.1080/17434440.2021.2013200.

Carrie, M. 2024. "Yale Medicine." Generative AI for Health Information: A Guide to Safe Use. https://www.yalemedicine.org/news/generative-ai-artificial-intelligence-for-health-info.

Chen, A. McDonald, Y. Zou, et al. (2022). "Trauma-Informed Computing: Towards Safer Technology Experiences for All." Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–20. https://doi.org/10.1145/3491102.3517475.

Clark, M., and S. Bailey. 2024. "Chatbots in Health Care: Connecting Patients to Information: Emerging Health Technologies." Canadian Agency for Drugs and Technologies in Health. http://www.ncbi.nlm.nih.gov/books/NBK602381/.

Coghlan, S., K. Leins, S. Sheldrick, M. Cheong, P. Gooding, and S. D'Alfonso. 2023. "To Chat or Bot to Chat: Ethical Issues With Using Chatbots in Mental Health." *Digital Health* 9: 20552076231183542. https://doi.org/10.1177/20552076231183542.

Collins, G. S., R. Whittle, G. S. Bullock, et al. 2024. "Open Science Practices Need Substantial Improvement in Prognostic Model Studies in Oncology Using Machine Learning." *Journal of Clinical Epidemiology* 165: 111199. https://doi.org/10.1016/j.jclinepi.2023.10.015.

Dhamala, J., T. Sun, V. Kumar, et al. (2021). "BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 862–872. https://doi.org/10.1145/3442188.3445924.

Duarte, E. F., and M. C. C. Baranauskas. 2016"Revisiting the Three HCIHCI Waves: A Preliminary Discussion on Philosophy of Science and

Research Paradigms." Proceedings of the 15th Brazilian Symposium on Human Factors in Computer Systems – IHC '16, 1–4. https://doi.org/10.1145/3033701.3033740.

Dupont, J. 1998. "The Concept of Trauma According to Ferenczi and Its Effects on Subsequent Psychoanalytical Research." *International Forum of Psychoanalysis* 7, no. 4: 235–241. https://doi.org/10.1080/080370698436736.

Eggleston, M., and L.-A. Noel. 2024. "Repairing the Harm of Digital Design Using a Trauma-Informed Approach." *Diseña* 24: 7. https://doi.org/10.7764/disena.24.Article.7.

Elyoseph, Z., D. Hadar-Shoval, K. Asraf, and M. Lvovsky. 2023. "ChatGPT Outperforms Humans in Emotional Awareness Evaluations." *Frontiers in Psychology* 14: 1199058. https://doi.org/10.3389/fpsyg.2023.1199058.

Fehr, J., B. Citro, R. Malpani, C. Lippert, and V. I. Madai. 2024. "A Trustworthy AI Reality-Check: The Lack of Transparency of Artificial Intelligence Products in Healthcare." *Frontiers in Digital Health* 6: 1267290. https://doi.org/10.3389/fdgth.2024.1267290.

Ferrara, E. 2024. "GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models." *Journal of Computational Social Science* 7: 549–569. https://doi.org/10.1007/s42001-024-00250-1.

Freed, D., J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N. Dell (2018). "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3173574.3174241.

Government of Canada. 2018. Defining Vulnerable Populations – Closed Consultation [Guidance], November 22, 2018. https://www.canada.ca/en/health-canada/services/chemical-substances/consulting-future-chemicals-management-canada/defining-vulnerable-populations.html.

Green, B. L., A. Murphy, and E. Robinson. 2024. "Accelerating Health Disparities Research With Artificial Intelligence." *Frontiers in Digital Health* 6: 1330160. https://doi.org/10.3389/fdgth.2024.1330160.

Greenberg, D. M., S. Baron-Cohen, N. Rosenberg, P. Fonagy, and P. J. Rentfrow. 2018. "Elevated Empathy in Adults Following Childhood Trauma." *PLoS One* 13, no. 10: e0203886. https://doi.org/10.1371/journal.pone.0203886.

Habib, H., Y. Zou, Y. Yao, et al. (2021). "Toggles, Dollar Signs, and Triangles: How to (In)Effectively Convey Privacy Choices With Icons and Link Texts." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–25. https://doi.org/10.1145/3411764.3445387.

Haimson, O. L., J. Buss, Z. Weinger, D. L. Starks, D. Gorrell, and B. S. Baron. 2020. "Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site." *Proceedings of the ACM on Human-Computer Interaction* 4, no. CSCW2: 1–27. https://doi.org/10.1145/3415195.

Hastings, J. 2024. "Preventing Harm from Non-Conscious Bias in Medical Generative AI." *Lancet. Digital Health* 6, no. 1: e2–e3. https://doi.org/10.1016/S2589-7500(23)00246-7.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33, no. 2–3: 61–83. https://doi.org/10.1017/S0140525X0999152X.

Henry, N., A. Flynn, and A. Powell. 2020. "Technology-Facilitated Domestic and Sexual Violence: A Review." *Violence Against Women* 26, no. 15–16: 1828–1854. https://doi.org/10.1177/1077801219875821.

Henry, N., and A. Powell. 2018. "Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research." *Trauma, Violence, & Abuse* 19, no. 2: 195–208. https://doi.org/10.1177/1524838016650189.

Holloway, I. W., S. Dunlap, H. E. del Pino, K. Hermanstyne, C. Pulsipher, and R. J. Landovitz. 2014. "Online Social Networking, Sexual Risk and Protective Behaviors: Considerations for Clinicians and Researchers." *Current Addiction Reports* 1: 220–228. https://doi.org/10.1007/s40429-014-0029-4.

Holmes, K. 2018. *Mismatch: How Inclusion Shapes Design*. The MIT Press. https://doi.org/10.7551/mitpress/11647.001.0001.

Hong, S., N. Lu, D. Wu, D. E. Jimenez, and R. L. Milanaik. 2020. "Digital Sextortion: Internet Predators and Pediatric Interventions." *Current Opinion in Pediatrics* 32, no. 1: 192–197. https://doi.org/10.1097/MOP.0000000000000854.

Houssami, N., G. Kirkpatrick-Jones, N. Noguchi, and C. I. Lee. 2019. "Artificial Intelligence (AI) for the Early Detection of Breast Cancer: A Scoping Review to Assess AI's Potential in Breast Screening Practice." *Expert Review of Medical Devices* 16: 351–362. https://www.tandfonline.com/doi/abs/10.1080/17434440.2019.1610387.

IBM Research. 2021, February 9. "What is generative AI? IBM Research." https://research.ibm.com/blog/what-is-generative-AI.

Kabudi, T., I. Pappas, and D. H. Olsen. 2021. "AI-Enabled Adaptive Learning Systems: A Systematic Mapping of the Literature." *Computers and Education: Artificial Intelligence* 2: 100017. https://doi.org/10.1016/j.caeai.2021.100017.

Kashima, Y. 2008. "A Social Psychology of Cultural Dynamics: Examining How Cultures Are Formed, Maintained, and Transformed." *Social and Personality Psychology Compass* 2, no. 1: 107–120. https://doi.org/10.1111/j.1751-9004.2007.00063.x.

Katims, I. 1995. "The Contrary Ideals of Individualism and Nursing Value of Care." *Scholarly Inquiry for Nursing Practice* 9, no. 3: 231–240; discussion 241-244.

Kelly, S., B. Lauren, and K. Nguyen (2021). "Trauma-informed Web Heuristics for Communication Designers." 39th ACM International Conference on the Design of Communication: Building Coalitions. Worldwide, SIGDOC 2021, October 12, 2021 – October 14, 2021, 172–176. https://doi.org/10.1145/3472714.3473638.

Knipscheer, J., M. Sleijpen, L. Frank, et al. 2020. "Prevalence of Potentially Traumatic Events, Other Life Events and Subsequent Reactions Indicative for Posttraumatic Stress Disorder in the Netherlands: A General Population Study Based on the Trauma Screening Questionnaire." *International Journal of Environmental Research and Public Health* 17, no. 5: 1725. https://doi.org/10.3390/ijerph17051725.

Kraft, E. 2006. "Cyberbullying: A Worldwide Trend of Misusing Technology to Harass Others." *Internet Society II: Advances in Education, Commerce & Governance* 1: 155–166. https://doi.org/10.2495/IS060161.

Lamensch, M. 2024. "Generative AI Tools Are Perpetuating Harmful Gender Stereotypes." *Centre for International Governance Innovation*. https://www.cigionline.org/articles/generative-ai-tools-are-perpetuating-harmful-gender-stereotypes/.

Larasati, R. (2023). "AI in Healthcare—Reflection on Potential Harms and Impacts (Short Paper)." In *CEUR Workshop Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI 2023)*, edited by P. K. Murukannaiah and T. Hirzle. Vol. 3456, 119–125. CEUR Workshop Proceedings (CEUR-WS.org). https://ceur-ws.org/Vol-3456/short3-2.pdf.

Larsson, S., and F. Heintz. 2020. "Transparency in Artificial Intelligence." *Internet Policy Review* 9, no. 2. https://policyreview.info/concepts/transparency-artificial-intelligence.

Lemonne, E. 2018, December 17. "Ethics Guidelines for Trustworthy AI." *Futurium – European Commission*. https://ec.europa.eu/futurium/en/ai-alliance-consultation.

Lucas, K. T. 2022. "Deepfakes and Domestic Violence: Perpetrating Intimate Partner Abuse Using Video Technology." *Victims & Offenders* 17, no. 5: 647–659. https://doi.org/10.1080/15564886.2022.2036656.

Machimbarrena, J. M., E. Calvete, L. Fernández-González, A. Álvarez-Bardón, L. Álvarez-Fernández, and J. González-Cabrera. 2018. "Internet

Risks: An Overview of Victimization in Cyberbullying, Cyber Dating Abuse, Sexting, Online Grooming and Problematic Internet Use." *International Journal of Environmental Research and Public Health* 15, no. 11: 2471. https://doi.org/10.3390/ijerph15112471.

MacLure, K., and A. Jones. 2021. "Domestic Abuse and Intimate Partner Violence: The Role of Digital by Design." *Journal of Adult Protection* 23, no. 5: 282–301.

Maras, M.-H., and A. Alexandrou. 2019. "Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos." *International Journal of Evidence & Proof* 23, no. 3: 255–262. https://doi.org/10.1177/1365712718807226.

Marr, B. 2024. *Building Trust in AI: The Case for Transparency*. Forbes. https://www.forbes.com/sites/bernardmarr/2024/05/03/building-trust-in-ai-the-case-for-transparency/.

Martínez, G., L. Watson, P. Reviriego, J. A. Hernández, M. Juarez, and R. Sarkar. 2024. "Epistemic Uncertainty in Artificial Intelligence." In *Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet*, edited by F. Cuzzolin and M. Sultana, 59–73. Switzerland: Springer Nature. https://doi.org/10.1007/978-3-031-57963-9_5.

Matt. 2024, June 5. "Trauma Informed Care Training in Virtual Reality | Workinman Interactive." https://workinman.com/vr-app-trauma-informed-care-training/.

Matthes, E. H. 2016. "Cultural Appropriation Without Cultural Essentialism?" *Social Theory and Practice* 42, no. 2: 343–366. https://www.jstor.org/stable/24871347.

Matthews, T., K. O'Leary, A. Turner, et al. (2017). "Stories From Survivors: Privacy & Security Practices When Coping With Intimate Partner Abuse." Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2189–2201. https://doi.org/10.1145/3025453.3025875.

McGowan, C. (2023). Generative AI and the Potential for Nefarious Use. https://www.isaca.org/resources/news-and-trends/industry-news/2023/generative-ai-and-the-potential-for-nefarious-use.

McMahon, H. V., and B. D. McMahon. 2024. "Automating Untruths: ChatGPT, Self-Managed Medication Abortion, and the Threat of Misinformation in a Post-Roe World." *Frontiers in Digital Health* 6: 1287186. https://doi.org/10.3389/fdgth.2024.1287186.

Meskó, B., and E. J. Topol. 2023. "The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare." *NPJ Digital Medicine* 6, no. 1: 120. https://doi.org/10.1038/s41746-023-00873-0.

Microsoft News Center. 2023. "Microsoft and Epic Expand Strategic Collaboration With Integration of Azure OpenAI Service | Epic." https://www.epic.com/epic/post/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service.

Musetti, A., V. Starcevic, V. Boursier, P. Corsano, J. Billieux, and A. Schimmenti. 2021. "Childhood Emotional Abuse and Problematic Social Networking Sites Use in a Sample of Italian Adolescents: The Mediating Role of Deficiencies in Self-Other Differentiation and Uncertain Reflective Functioning." *Journal of Clinical Psychology* 77, no. 7: 1666–1684. https://doi.org/10.1002/jclp.23138.

Nadeem, M., A. Bethke, and S. Reddy. 2021 "StereoSet: Measuring Stereotypical Bias in Pretrained Language Models." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, edited by C. Zong, F. Xia, W. Li and R. Navigli (Volume 1: Long Papers), 5356–5371. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.416.

Newland, R., M. Lawrence, S. Tyndall, and J. Waterall. 2022. "Vulnerability and Trauma-Informed Practice: What Nurses Need to Know." *British Journal of Nursing* 31, no. 12: 660–662. https://doi.org/10.12968/bjon.2022.31.12.660.

Ngũnjiri, A., P. Memiah, R. Kimathi, et al. 2023. "Utilizing User Preferences in Designing the AGILE (Accelerating Access to Gender-Based Violence Information and Services Leveraging on Technology Enhanced) Chatbot." *International Journal of Environmental Research and Public Health* 20, no. 21: 7018. https://doi.org/10.3390/ijerph20217018.

Nicki, A. 2001. "The Abused Mind: Feminist Theory, Psychiatric Disability, and Trauma." *Hypatia* 16, no. 4: 80–104. https://doi.org/10.1111/j.1527-2001.2001.tb00754.x.

Nsiah, C., M. Siakwa, and J. P. K. Ninnoni. 2019. "Registered Nurses' Description of Patient Advocacy in the Clinical Setting." *Nursing Open* 6, no. 3: 1124–1132. https://doi.org/10.1002/nop2.307.

Nugent, N. R., J. A. Sumner, and A. B. Amstadter. 2014. "Resilience After Trauma: From Surviving to Thriving." *European Journal of Psychotraumatology* 5, no. 1: 25339. https://doi.org/10.3402/ejpt.v5.25339.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464: 447–453. https://doi.org/10.1126/science.aax2342.

Ouyang, L., J. Wu, X. Jiang, et al. (2024). "Training Language Models to Follow Instructions With Human Feedback." Proceedings of the 36th International Conference on Neural Information Processing Systems, 27730–27744.

Raza, M. M., K. P. Venkatesh, and J. C. Kvedar. 2024. "Generative AI and Large Language Models in Health Care: Pathways to Implementation." *NPJ Digital Medicine* 7, no. 1: 62. https://doi.org/10.1038/s41746-023-00988-4.

Reddy, S. 2024. "Generative AI in Healthcare: An Implementation Science Informed Translational Path on Application, Integration and Governance." *Implementation Science* 19, no. 1: 27. https://doi.org/10.1186/s13012-024-01357-9.

Regehr, K., A. Birze, and C. Regehr. 2022. "Technology Facilitated Re-Victimization: How Video Evidence of Sexual Violence Contributes to Mediated Cycles of Abuse." *Crime, Media, Culture: An International Journal* 18, no. 4: 597–615. https://doi.org/10.1177/17416590211050333.

Responsible AI Institute. 2020. "Accelerating Responsible AI Adoption." *Responsible AI*. https://www.responsible.ai/.

Rey Velasco, E., Z. Demjén, and T. C. Skinner. 2024. "Digital Empathy in Behaviour Change Interventions: A Survey Study on Health Coach Responses to Patient Cues." *Digital Health* 10: 20552076231225889. https://doi.org/10.1177/20552076231225889.

Rodriguez, D. V., K. Lawrence, J. Gonzalez, et al. 2024. "Leveraging Generative AI Tools to Support the Development of Digital Solutions in Health Care Research: Case Study." *JMIR Human Factors* 11: e52885. https://doi.org/10.2196/52885.

Rousay Victoria. 2023. "Sexual Deepfakes and Image-Based Sexual Abuse: Victim-Survivor Experiences and Embodied Harms." Master's thesis, Harvard University. https://www.proquest.com/openview/d44f0222387ea96c8a20fe9f517c8350/1?pq-origsite=gscholar&cbl=18750&diss=y.

Schüz, B., and M. Urban. 2020. "Unerwünschte Effekte Digitaler Gesundheitstechnologien: Eine Public-Health-Perspektive." *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz* 63, no. 2: 192–198. https://doi.org/10.1007/s00103-019-03088-5.

Seyyed-Kalantari, L., H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi. 2021. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-Served Patient Populations." *Nature Medicine* 27, no. 12: 2176–2182. https://doi.org/10.1038/s41591-021-01595-0.

Shanklin, R., M. Samorani, S. Harris, and M. A. Santoro. 2022. "Ethical Redress of Racial Inequities in AI: Lessons From Decoupling Machine Learning From Optimization in Medical Appointment Scheduling." *Philosophy & Technology* 35, no. 4: 96. https://doi.org/10.1007/s13347-022-00590-8.

Shaughnessy, K., C. J. Fehr, M. Ashley, et al. 2022. "Technology-Mediated Sexual Interactions, Social Anxiety, and Sexual Wellbeing: A Scoping Review." *European Journal of Investigation in Health, Psychology and Education* 12, no. 8: 904–932. https://doi.org/10.3390/ejihpe12080066.

Sheikh, M., and M. M. Rogers. 2024. "Technology-Facilitated Sexual Violence and Abuse in Low and Middle-Income Countries: A Scoping Review." *Trauma, Violence & Abuse* 25, no. 2: 1614–1629. https://doi.org/10.1177/15248380231191189.

Simko, L., A. Lerner, S. Ibtasam, F. Roesner, and T. Kohno. 2018. "Computer Security and Privacy for Refugees in the United States." In *2018 IEEE Symposium on Security and Privacy (SP)*, 409–423. https://doi.org/10.1109/SP.2018.00023.

Skornyakov, Anton. 2023. "Using Trauma-Informed Approaches in Agile Environments." *InfoQ*. https://www.infoq.com/articles/trauma-informed-agile/.

Stonard, K. E., E. Bowen, K. Walker, and S. A. Price. 2017. "They'll Always Find a Way to Get to You": Technology Use in Adolescent Romantic Relationships and Its Role in Dating Violence and Abuse." *Journal of Interpersonal Violence* 32, no. 14: 2083–2117. https://doi.org/10.1177/0886260515590787.

Straw, I., and L. Tanczer. 2023. "Safeguarding Patients From Technology-Facilitated Abuse in Clinical Settings: A Narrative Review." *PLoS Digital Health* 2, no. 1: e0000089. https://doi.org/10.1371/journal.pdig.0000089.

Tebes, J. K., R. B. Champine, S. L. Matlin, and M. J. Strambler. 2019. "Population Health and Trauma-Informed Practice: Implications for Programs, Systems, and Policies." *American Journal of Community Psychology* 64, no. 3–4: 494–508. https://doi.org/10.1002/ajcp.12382.

Tengkawan, J., R. Agnihotri, and R. S. Minhas. 2022. "Creating Inclusive Digital Health Resources for Marginalised Culturally Diverse Families: A Call to Action." *BMJ Paediatrics Open* 6, no. 1: e001626. https://doi.org/10.1136/bmjpo-2022-001626.

UK Council for Internet Safety. 2019. "Adult Online Hate, Harassment and Abuse: A rapid evidence assessment." Gov.Uk. https://www.gov.uk/government/publications/adult-online-hate-harassment-and-abuse-a-rapid-evidence-assessment.

Ulnicane, I. 2024. "Intersectionality in Artificial Intelligence: Framing Concerns and Recommendations for Action." *Social Inclusion* 12: 7543. https://doi.org/10.17645/si.7543.

Venkatasubramanian, K., and T.-M. Ranalli (2022). "Designing Post-Trauma Self-Regulation Apps for People With Intellectual and Developmental Disabilities." Proceedings of the 24th International Acm Sigaccess Conference on Computers and Accessibility, 1–14. https://doi.org/10.1145/3517428.3544798.

Wang, Y., and M. Kosinski. 2018. "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images." *Journal of Personality and Social Psychology* 114, no. 2: 246–257. https://doi.org/10.1037/pspa0000098.

WHO. (2024). "WHO Unveils a Digital Health Promoter Harnessing Generative AI for Public Health." https://www.who.int/news/item/02-04-2024-who-unveils-a-digital-health-promoter-harnessing-generative-ai-for-public-health.

Wong, K. L. Y., A. Sixsmith, L. Remund, M. Pomati, A. Jolly, and J. Rees. 2021. "Older Adults' Access to Information and Referral Services Using Technology in British Columbia, Canada: Past Learnings and Learnings Since COVID-19." In *Social Policy Review*. (Vol.33)., 161–180. Policy Press. https://bristoluniversitypressdigital.com/display/book/9781447359739/ch008.xml.

Wornow, M., Y. Xu, R. Thapa, et al. 2023. "The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records." *NPJ Digital Medicine* 6, no. 1: 135. https://doi.org/10.1038/s41746-023-00879-8.

Wu, E., K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou. 2021. "How Medical Ai Devices Are Evaluated: Limitations and Recommendations From an Analysis of Fda Approvals." *Nature Medicine* 27, no. 4: 582–584. https://doi.org/10.1038/s41591-021-01312-x.

Wynants, L., B. Van Calster, G. S. Collins, et al. 2020. "Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal." *BMJ* 369: m1328. https://doi.org/10.1136/bmj.m1328.

Xu, D., S. Fan, and M. Kankanhalli (2023). "Combating Misinformation in the Era of Generative AI Models." Proceedings of the 31st ACM International Conference on Multimedia, 9291–9298. https://doi.org/10.1145/3581783.3612704.

Xu, Y., N. Bradford, and R. Garg. 2023. "Transparency Enhances Positive Perceptions of Social Artificial Intelligence." *Human Behavior and Emerging Technologies* 2023, no. 1: 5550418. https://doi.org/10.1155/2023/5550418.

Zack, T., E. Lehman, M. Suzgun, et al. 2024. "Assessing the Potential of GPT-4 to Perpetuate Racial and Gender Biases in Health Care: A Model Evaluation Study." *The Lancet Digital Health* 6, no. 1: e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X.

Ziebland, S., E. Hyde, and J. Powell. 2021. "Power, Paradox and Pessimism: On the Unintended Consequences of Digital Health Technologies in Primary Care." *Social Science & Medicine (1982)* 289: 114419. https://doi.org/10.1016/j.socscimed.2021.114419.