PERSPECTIVE OPEN

Check for updates

Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare

David Oniani[1], Jordan Hilsman[1], Yifan Peng[2], Ronald K. Poropatich[3,4], Jeremy C. Pamplin[5], Gary L. Legault[6,7] and Yanshan Wang[1,8,9,10,11] ✉

In 2020, the U.S. Department of Defense officially disclosed a set of ethical principles to guide the use of Artificial Intelligence (AI) technologies on future battlefields. Despite stark differences, there are core similarities between the military and medical service. Warriors on battlefields often face life-altering circumstances that require quick decision-making. Medical providers experience similar challenges in a rapidly changing healthcare environment, such as in the emergency department or during surgery treating a life-threatening condition. Generative AI, an emerging technology designed to efficiently generate valuable information, holds great promise. As computing power becomes more accessible and the abundance of health data, such as electronic health records, electrocardiograms, and medical images, increases, it is inevitable that healthcare will be revolutionized by this technology. Recently, generative AI has garnered a lot of attention in the medical research community, leading to debates about its application in the healthcare sector, mainly due to concerns about transparency and related issues. Meanwhile, questions around the potential exacerbation of health disparities due to modeling biases have raised notable ethical concerns regarding the use of this technology in healthcare. However, the ethical principles for generative AI in healthcare have been understudied. As a result, there are no clear solutions to address ethical concerns, and decision-makers often neglect to consider the significance of ethical principles before implementing generative AI in clinical practice. In an attempt to address these issues, we explore ethical principles from the military perspective and propose the “GREAT PLEA” ethical principles, namely Governability, Reliability, Equity, Accountability, Traceability, Privacy, Lawfulness, Empathy, and Autonomy for generative AI in healthcare. Furthermore, we introduce a framework for adopting and expanding these ethical principles in a practical way that has been useful in the military and can be applied to healthcare for generative AI, based on contrasting their ethical concerns and risks. Ultimately, we aim to proactively address the ethical dilemmas and challenges posed by the integration of generative AI into healthcare practice.

npj Digital Medicine (2023)6:225 ; https://doi.org/10.1038/s41746-023-00965-x

INTRODUCTION

Artificial Intelligence (AI) plays an ever-increasing role in our daily lives and has influenced fields from online advertising to sales and from the military to healthcare. With the ongoing AI arms race in the Russia-Ukraine War, it is expected that AI-powered lethal weapon systems will become commonplace in warfare[1]. Although AI has shown promise in numerous successful applications, there remains a pressing need to address ethical concerns associated with these applications. There are dire consequences if an AI system selects an incorrect target potentially killing non-combatants or friendly forces. Seeing the rapid emergence of AI and its applications in the military, the United States Department of Defense (DOD) disclosed ethical principles for AI in 2020[2]. This document emphasized five core principles, aiming for responsible, equitable, traceable, reliable, and governable AI[2]. In addition, the North Atlantic Treaty Organization (NATO) also released principles for the use of AI in military, including lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation[3]. The success of these ethical principles has also been demonstrated through their ability to adopt and embed AI mindfully, taking into account AI’s potential dangers, which the Pentagon is determined to avoid[4]. Clearly, prominent military organizations demonstrate a cautious approach toward adopting AI and are actively implementing measures to mitigate the risks associated with its potential malicious uses and applications.

On the other hand, AI has had a direct impact on the healthcare industry, with discussions ranging from the uses of AI as an assistant to medical personnel[5–7] to AI replacing entire clinical departments[8,9]. The use and impact of AI in clinical Natural Language Processing (NLP) in the context of Electronic Health Records (EHRs) have been profound[10–13]. Similar to military organizations, the World Health Organization (WHO) has also released a document discussing the ethics and governance of AI for health[14].

Generative AI, as the name suggests, refers to AI techniques that can be used to create or produce various types of new contents, including text, images, audio, and videos. The rate of development of generative AI has been staggering, with many industries and researchers finding its use in fields such as finance[15], collaborative writing[16], email communication[17], and cyber threat intelligence[18]. Generative AI has also become an active area of research in the healthcare domain[19,20], with

[1]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA. [2]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [3]Division of Pulmonary, Allergy, Critical Care & Sleep Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [4]Center for Military Medicine Research, University of Pittsburgh, Pittsburgh, PA, USA. [5]Telemedicine & Advanced Technology Research Center, US Army, Fort Detrick, Frederick, MD, USA. [6]Department of Surgery, Uniformed Services University, Bethesda, MD, USA. [7]Virtual Medical Center, Brooke Army Medical Center, San Antonio, TX, USA. [8]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA. [9]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA. [10]Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, USA. [11]University of Pittsburgh Medical Center, Pittsburgh, PA, USA. ✉email: yanshan.wang@pitt.edu

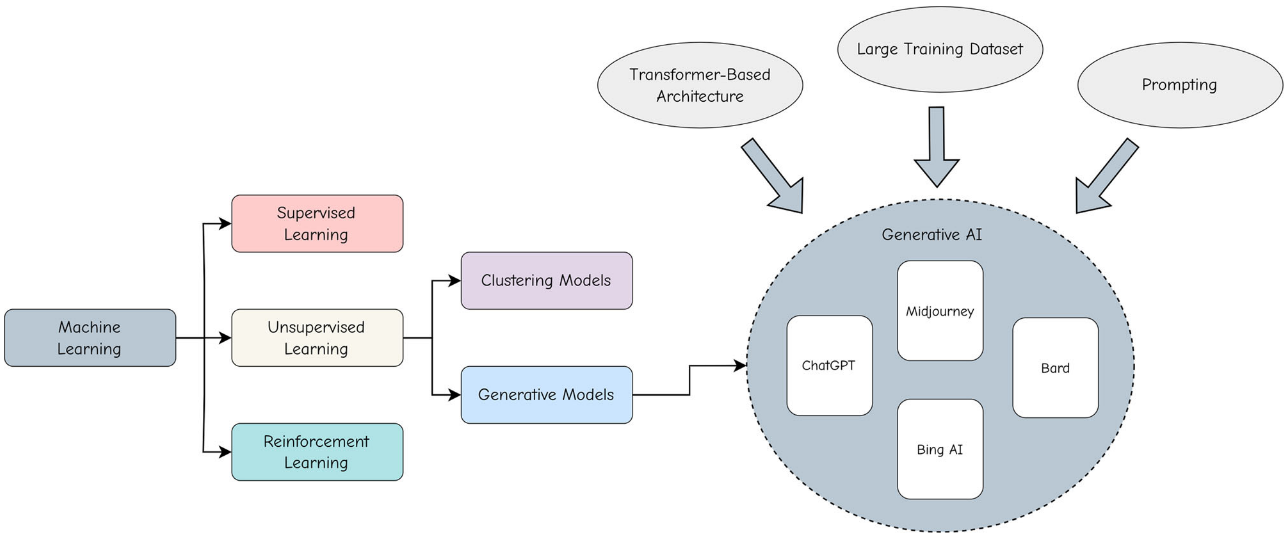


Fig. 1 The relationship between general ML and modern generative AI. The figure provides an overview of ML subfields, establishes relationships among these subfields, and shows the path to generative AI.

applications such as clinical documentation[21] and evidence-based medicine summarization[22].

Despite many successful and promising AI applications, ethics has been one of the more controversial subjects of discussion in the AI community, with diverging views and a plethora of opinions[23,24]. Ethics deals with how one decides what is morally right or wrong and is one of the pivotal aspects that we, as the AI research community, have to consider carefully. Given the recent emergence of generative AI models and their initial enthusiasm in healthcare, our community must seriously consider ethical principles before integrating these techniques into practical use. The military and healthcare are notably similar in many ways, such as organizational structure, high levels of stress and risk, decision-making processes, reliance on protocols, and dominion over life and death. Given these parallels, successful implementation of ethical principles in military applications, and the lack of specific solutions to generative AI ethics in healthcare, we propose to adopt and expand ethical principles, from military to healthcare, to govern the application of generative AI in healthcare applications.

WHAT IS GENERATIVE ARTIFICIAL INTELLIGENCE?

Generative AI refers to AI that is used primarily for generating data, often in the form of audio, text, and images. However, in this manuscript, we choose not to follow such a general definition and instead, focus on a particular type of generative AI. In this section, we describe "modern" generative AI, discuss why it is important, and compare it to the term that has become so popular—"AI."

Modern AI is dominated by Machine Learning (ML) methods, which leverage statistical algorithms and large amounts of data to gradually improve model performance. ML methods could roughly be classified into supervised, unsupervised, and reinforcement learning (Fig. 1). Supervised ML relies on labeled input (supervision), while unsupervised learning needs no human supervision. Reinforcement learning takes a different approach and, instead, attempts to design intelligent agents by rewarding desired behaviors and punishing undesired ones. Popular generative AI models are typically pre-trained in an unsupervised manner.

The pre-trained generative AI models could generate novel and diverse outputs, including, but not limited to, text, images, audio, or videos. Recently, the most popular generative AI model for language generation is ChatGPT[25], which was reported to have an estimated 100 million monthly active users in January 2023[26]. The model architectures for ChatGPT, previously known as GPT-3.5[27], and more recent GPT-4[28], are built upon the design principles of its GPT[29] (Generative Pre-trained Transformer) predecessors, GPT-2[30] and GPT-3[31]. Many state-of-the-art generative AI models, also known as Large Language Models (LLMs), share a similar transformer-based architecture[32].

The well-known generative AI models used for image generation from text prompts, such as Stable Diffusion[33] and DALL-E 2[34], employ a combination of the diffusion process[35] and a transformer-based architecture similar to the one used in GPT models. All of the models are characterized by unsupervised training on very large datasets[36]. The same is true of models that generate videos.

Most of these generative AI models also rely on a method called prompting[37], which lets users input a natural language description of a task and uses it as a context to generate useful information. This process is also sometimes referred to as in-context learning.

When referring to "modern" generative AI or simply generative AI, we are describing a transformer-based machine learning model trained in an unsupervised manner on extensive datasets and specifically optimized for generating valuable data through prompts. This description also aligns harmoniously with existing research and studies[38–40].

While generative AI shows promising results, dangerous outcomes in healthcare can arise from a number of issues, including:

- Algorithmic bias[41,42]
- Hallucination[43,44]
- Poor commonsense reasoning[44,45]
- Lack of generally agreed model evaluation metrics[46,47]

All of these issues are common for generative AI in general, but more so in the healthcare domain, where algorithmic bias may result in the mistreatment of patients[48], hallucination may carry misinformation[49], poor commonsense reasoning can result in confusing interactions[50], and lack of general and domain-specific metrics can make it difficult to validate the robustness of the AI system[51]. Furthermore, in the context of healthcare, there are concerns about leaking Protected Health Information (PHI)[52] as well as lacking empathy to patients[53].

Such concerns can also be present in other forms of AI, but given the practical differences present in generative AI, the risks become elevated. First, due to the interactive nature of generative

AI, often paired with the ability to hold human-like dialogs (e.g., ChatGPT), it can make misinformation sound convincing. Second, since generative AI models combine various sources of large-scale data[36], the risk of training on biased data sources increases. Third, the standard evaluation metrics, such as precision, recall, and F1 score, become difficult to use and are less likely to reflect human judgment[47]. Finally, due to its ease of use, generative AI can be widely adopted in many fields and domains of healthcare[49], which naturally increases the aforementioned risks.

Overall, the importance of ethical considerations for generative AI in healthcare cannot be understated. From the human-centered perspective, the ultimate goal of generative AI is to enhance and augment human's creativity, productivity, and problem-solving capabilities, which is well aligned with the goal of healthcare in improving patient care. If the generative AI system is not used ethically and does not reflect our values, its role as a tool for improving the lives of people will greatly diminish.

AI APPLICATIONS IN MILITARY VS. HEALTHCARE

With the increasing prevalence of AI, it has been in the best interest of military organizations to understand and integrate AI into their operations and strategies to be at the cutting edge of security and technology in conflict or emergency. Various military AI technologies for generative purposes have also been developed, including Intelligent Decision Support Systems (IDSSs) and Aided Target Recognition (AiTC), which assist in decision-making, target recognition, and casualty care in the field[54–56]. Each of these uses of AI in military operations reduces the mental load of operators in the field and helps them take action more quickly. Just as military uses of AI can save lives on the battlefield, AI can help save lives by assisting clinicians in diagnosing diseases and reducing risks to patient safety[57–59]. Uses of generative AI in healthcare help improve the efficiency of professionals caring for patients. Applications of generative AI in healthcare include medical chatbots, disease prediction, CT image reconstruction, and clinical decision support tools[60–63]. The benefits of such uses are two-fold, in that they can help healthcare professionals deliver a higher level of care to their patients, as well as improve the workload within clinics and hospitals.

People may question that developing AI models for military and healthcare purposes hinges on distinct ideological underpinnings reflecting unique priorities. In the military context, AI models are primarily designed to enhance the efficiency, precision, and strategic capabilities of both defensive and offensive operations. The focus is on applications such as surveillance, target recognition, cyber defense, autonomous weaponry, and battlefield analytics. Potential future uses of AI for offensive actions such as coordinating drone attacks may oppose any healthcare principle, yet is vital for the military strategy. The fundamental ideological perspective here is the protection of national security interests, force multiplication, and minimizing human risk in conflict zones.

On the other hand, the use of AI in healthcare is driven by the principles of enhancing patient care, improving health outcomes, and optimizing the efficiency of healthcare systems. The development of AI models in this sector aims to personalize treatments, improve diagnostic accuracy, predict disease progression, and streamline administrative tasks, among other uses. The central ideology is the betterment of human health and well-being. While we acknowledge the different ideological foundations in military and healthcare due to the contrasting objectives, we argue that both military and healthcare sectors illustrate a compelling convergence of priorities for the applications of AI.

Specifically, their shared focus on application validity, attention to practical implementation, and prioritization of a human-centered approach have emerged as significant commonalities. First, concerning application validity, both fields recognize the crucial importance of robust, reliable AI systems. These systems need to function accurately and rapidly under diverse, often challenging, conditions to fulfill their designated tasks, whether it identifies potential security threats in a complex battlefield or detects subtle abnormalities in medical images. Second, there is an evident emphasis on implementation. Beyond the theoretical development of AI models, the critical question for both sectors centers around how these models can be effectively incorporated into real-world systems, often involving multiple human and technological stakeholders. Finally, a human-centered perspective is paramount. This means ensuring that AI technologies augment, rather than replace, human decision-making capacities and are employed in ways minimizing potential harm. In healthcare, this involves developing AI applications that can improve patient outcomes and experience while supporting healthcare providers in their work. Thus, these three factors represent key shared priorities in the utilization of AI across military and healthcare contexts.

AI has been seamlessly woven into the military's technology fabric for several decades, serving as the backbone for various advancements ranging from autonomous drone weapons to intelligent cruise missiles[64,65]. The track record of robust results and reliable outcomes in complex and high-risk environments implicitly engage with foundational ethical principles. The ethical guidelines established from military AI implementations have provided a road map for the incorporation of AI in healthcare scenarios. However, the integration of AI is relatively new to the healthcare sector, let alone generative AI, and ethical principles are neither widely implemented nor specifically designed for generative AI. While healthcare has begun to adopt generative AI technologies more recently[66], there are immense opportunities for this field to glean ethical insights from the history of military application.

IDENTIFYING ETHICAL CONCERNS AND RISKS

A RAND Corporation study raised various concerns about the use of AI in warfare, shown in Figure 2.3 of the research report[67]. These concerns fall into the following categories: increasing risk of war, increased errors, and misplaced faith in AI. Although AI can allow personnel to make decisions and strategies more quickly, some experts consider this a downside, as actions taken without proper consideration could have serious repercussions, like increasing the risk of war[67]. International standards for warfare like the Law of Armed Conflict (LOAC) and Geneva Conventions lay out guidelines for target identification specifying that attacks must first distinguish between combatant and noncombatant targets before taking action to minimize harm to civilians[68,69]. Because combatants are not always identifiable visually, some claim that reading body language to differentiate a civilian from a combatant necessitates a Human-In-The-Loop (HITL) decision-making process[70].

Maintaining data privacy for users of generative AI technologies is critical, as both patient data and military data are highly sensitive, and would be damaging if leaked[71]. If an AI implementation collects PHI, it should be secure against breaches, and any disclosures of this protected data must comply with Health Insurance Portability and Accountability Act (HIPAA) guidelines[72]. These implementations must experience few errors as healthcare is a safety-critical domain where the patient harm is unacceptable[73], and errors in these systems or algorithms could cause more harm than any physician would be capable of, as many hospitals and clinics would be using the same systems and experiencing the same errors[74].

Additional concerns present in the military and healthcare are trust between humans and AI and the lack of accountability. When there exists human-and-AI collaboration to perform a task, trust must be optimal, as shown in Fig. 2. Too much trust in AI systems can lead to overuse of the AI when it is not in the best interest of

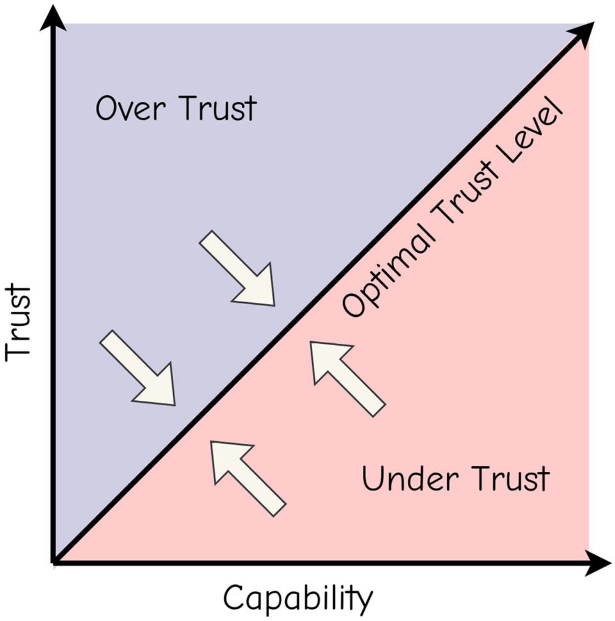


Fig. 2 Optimization of trust in AI. The figure depicts the relationship between trust and capability. Too much trust in AI systems can lead to overuse of the AI when it is not in the best interest of patients or operators, and too little trust can lead to underuse of the system when it would be better to use it.

patients or operators[75], and too little trust can lead to underuse of the system when it would be better to use it[76]. In both situations, the root cause is operators not knowing the capabilities and limitations of the systems they interact with[77]. Misuse can lead to non-typical errors, such as fratricide in the military or patient harm in a hospital[78,79]. While the AI must be transparent in its decision-making, the use of AI must be accompanied by sufficient education on the use and limitations of AI systems so that operators are less likely to make dangerous errors. A lack of accountability can possibly arise in military or healthcare use of generative AI because military operators or clinicians do not have direct control over the actions determined by the AI.

In the same research report by RAND Corporation[67], authors showed (Figures 7.8, 7.9 in the report) that the general public views autonomous systems taking military action with human authorization favorably while strongly disagreeing with combat action without human authorization. The parallel can be drawn with healthcare, where patients express concerns over the use of AI for medical purposes without human (e.g., physician, nurse, etc.) involvement[80]. These results could be due to the perceived lack of accountability, which is considered something that could entirely negate the value of AI, as a fully autonomous system that makes its own decisions distances military operators or clinicians from the responsibility of the system's actions[81]. In healthcare, it is critical that the systems are transparent due to their proximity to human lives and that patients understand how clinicians use these recommendations. The burden of accountability in the healthcare sector falls to both the clinicians and the developers of the AI systems, as the decisions made are a product of the algorithm, and the use of these recommendations falls to the clinicians[82].

Finally, ethical concerns of equity, autonomy, and privacy regarding the use of generative AI must also be considered. In healthcare settings, biased algorithms or biased practices can lead to certain patient groups receiving lower levels of care[83]. Biased outcomes could be due to biased algorithms, poor data collection, or a lack of diversity[84]. There must be minimal bias in developing AI systems in healthcare, both in the algorithm and the data used for training. Furthermore, if known, the sources of bias must also be disclosed to ensure transparency and prevent inappropriate use. The issue of human autonomy when developing generative AI is especially pertinent in healthcare, as both patient and clinician autonomy must be respected[85]. It is crucial that a framework is accepted to prevent any data breaches and ensure security measures are up to date and robust.

These risks and ethical concerns surrounding generative AI in military and healthcare applications necessitate principles for the ethical use of AI. One of the earliest sets of principles published for responsible development and use of AI comes from Google, who did so in response to their employees petitioning their CEO as they disagreed with Google working with the DOD on Project Maven to assist in identifying objects in drone images[86,87] in 2018. These principles outline how Google will develop AI responsibly and state what technologies they will not create, like those that cause harm or injure people, provide surveillance that violates international policies, and any technologies that go against international law and human rights[88]. By examining the differences and similarities between risks and ethical concerns in military and healthcare applications of generative AI, we can establish guiding principles for the responsible development and use of generative AI in healthcare.

GREAT PLEA ETHICAL PRINCIPLES FOR GENERATIVE AI IN HEALTHCARE

As AI usage has spread throughout the military and other fields, many organizations have recognized the necessity of articulating their ethical principles and outlining the responsibilities associated with applying AI to their operations. There are several ethical principles for AI published by various organizations, including the U.S. Department of Defense[2], NATO[3], the American Medical Association (AMA)[89], the World Health Organization (WHO)[14], and the Coalition for Health AI (CHAI)[90]. The AI ethical principles for DOD and NATO are similar, with NATO having an added focus on adherence to international law. For the development of AI for healthcare, the WHO has published its own ethical principles, including protecting human autonomy, human well-being and safety, transparency and explainability, responsibility and accountability, inclusiveness, and responsive development. Similarly, the AMA promotes AI systems that should be user-centered, transparent, reproducible, avoid exacerbating healthcare disparities, and safeguard the privacy interests of patients and other individuals. Finally, there is the Blueprint for an AI Bill of Rights published by the U.S. Office of Science and Technology Policy (OSTP)[91], which has provisions for AI systems to be safe and effective, protected against algorithmic discrimination, protect user data, have accessible documentation, and offer human alternatives.

Among the various sets of principles, we see common themes such as accountability and human presence. The DOD and NATO both emphasize the importance of integrating human responsibility into the development and life cycle of an AI system, as well as ensuring these systems are governable to address errors that may arise during use. The AMA and WHO policies both highlight a human-centered design philosophy protecting human autonomy and explicitly mention the need for inclusiveness and equity in the healthcare use of AI to prevent care disparity. These principles each provide unique perspectives for developing AI for healthcare use. However, no set of principles encompasses all ethical concerns that healthcare providers or patients may have[92]. Adopting the principles of the DOD and NATO is advantageous due to each principle's practical definition. These principles are outlined with a focus on what actions can be taken by personnel developing AI systems, and how end-users would interact with the systems.

The existing principles establish a good foundation for the ethical development and utilization of AI in healthcare. However,

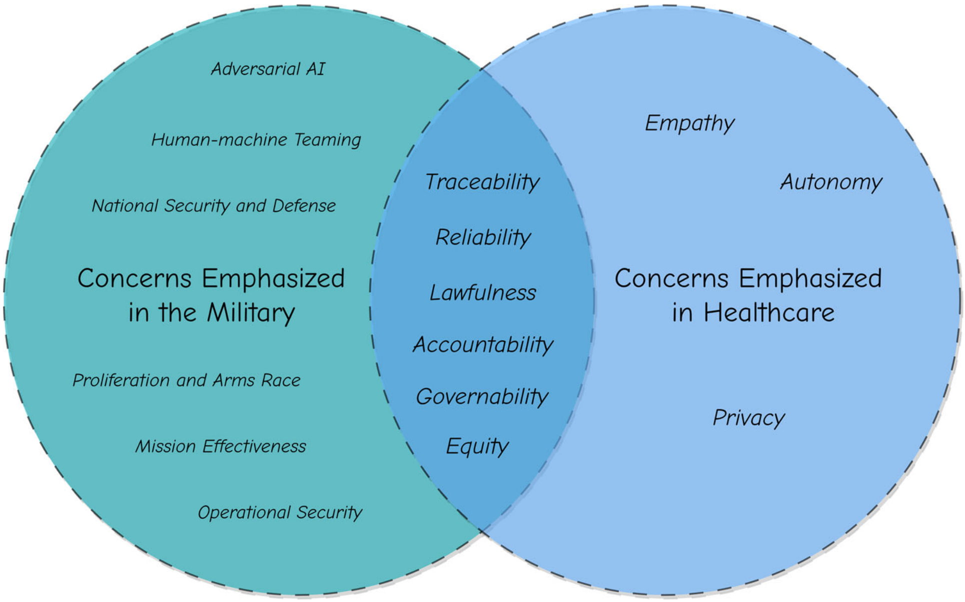


Fig. 3 Adoption and expansion of existing ethical principles from military to healthcare. The figure illustrates the commonalities and differences in ethical principles between military and healthcare. In our assessment, traceability, reliability, lawfulness, accountability, governability, and equity are the ethical principles that both fields have in common. At the same time, ethical principles, such as empathy and privacy, are emphasized in healthcare, whereas ethical principles, such as national security and defense, are emphasized in the military.

action must be taken to tailor these principles for generative AI. By examining the risks and concerns surrounding the use of generative AI in healthcare, comparing them to the risks and concerns of generative AI in the military, and by expanding these principles, we can have a set of principles that fulfill our needs[93]. Therefore, we use DOD and NATO guidelines as the starting point for the set of ethical principles, and expand them to meet the needs in healthcare. The expansion is done by incorporating principles that support the betterment of mankind rather than defeating adversaries.

Figure 3 shows the framework that we used for adopting and expanding ethical principles, established by various organizations, for the healthcare applications of AI. Where similarities are present in the concerns between military and healthcare use of generative AI, it is possible to adopt principles for use, such as traceability, reliability, lawfulness, accountability, governability, and equity. In instances when healthcare has unique circumstances or requires additional nuance, the principles related to those matters must be expanded to fit into the world of medicine, such as empathy, autonomy, and privacy. There are many concerns specific to the military that are unsuitable for forming ethical principles in healthcare, such as national security and defense, mission effectiveness, operational security, adversarial AI[94], human-machine teaming, rules of engagement, rapid deployment and adaptation, and proliferation and arms race. Figure 3 also shows some of these concerns (a non-exhaustive list), which we included to highlight that the adoption or expansion of principles must be based on shared concerns. Furthermore, we want to emphasize the need for having safeguards and methods to detect and mitigate military-specific properties of AI deployed in healthcare settings, by including governability, accountability, and traceability.

A detailed mapping of the proposed ethical principles to those used by DOD, NATO, and WHO guidelines is shown in Table 1. As shown in the table, all principles, except for privacy, empathy, and autonomy, directly align with either DOD or NATO guidelines. In cases where the principle indirectly aligns with our proposed principles, Table 1 uses a star (*) prefix. As for privacy, empathy, and autonomy, despite not being related to the ethical principles in military organizations (i.e., DOD and NATO), WHO guidelines directly or indirectly align with all three. Their inclusion was also due to the quality of betterment of mankind and mitigation of concerns specific to healthcare.

In summary, we propose the "GREAT PLEA" ethical principles for generative AI in healthcare, namely Governability, Reliability, Equity, Accountability, Traceability, Privacy, Lawfulness, Empathy, and Autonomy. The GREAT PLEA ethical principles demonstrate our great plea for the community to prioritize these ethical principles when implementing and utilizing generative AI in practical healthcare settings. Fig. 4 shows the summary cards for the GREAT PLEA ethical principles. In the following, we will delve into a comprehensive explanation of each individual principle.

Governability

Governability is the ability of a system to integrate processes and tools which promote and maintain its capability and ensure meaningful human control[95]. Standards for the governability of AI systems, as established by the DOD and NATO, emphasize the importance of ensuring that while AI systems fulfill their intended functions, humans must retain the ability to identify and prevent unintended consequences. In the event of any unintended behavior, human intervention to disengage or deactivate the deployed AI system should be possible. These standards can be adopted for the use of generative AI in healthcare. Due to the potential of widespread implementation of generative AI systems, where numerous hospitals may be using the same systems, these standards must be considered[74]. Suppose a generative AI system, deployed across multiple clinics, poses a risk of harm to a patient. In that case, it is crucial to recognize that numerous patients across clinics could be vulnerable to the same error. Risk to patients amplifies as healthcare expands to patient homes with remote patient monitoring or with online tools outside the clinic. Ideally, humans, whether they develop or implement the system, should possess the capability to deactivate it without disrupting the regular patient care activities in the clinics. There must be explicit guidelines for monitoring generative AI systems for potential errors, deactivation to prevent more damage when an error occurs, remedying errors, and interaction to reduce operator errors. With these guidelines in place, personnel in charge of the

Table 1. Alignment of GREAT PLEA ethical principles with DOD, NATO, and WHO guidelines.

Principle	DOD	NATO	WHO
Governability	Governable	Governability	*Promote AI that is responsive and sustainable
Reliability	Reliable	Reliability	*Promote human well-being, human safety, and the public interest
Equity	Equitable	Bias mitigation	Ensure inclusiveness and equity
Accountability	*Responsible	Responsibility and accountability	Foster responsibility and accountability
Traceability	Traceable	Explainability and traceability	Ensure transparency, explainability, and intelligibility
Privacy	None	None	*Ensure transparency, explainability, and intelligibility
Lawfulness	None	Lawfulness	*Promote AI that is responsive and sustainable
Empathy	None	None	*Promote human well-being, human safety, and the public interest
Autonomy	None	None	Protect autonomy

*mark: indirectly aligned principle

system can quickly be notified of any unintended behavior and respond quickly and appropriately.

Reliability

Reliability is the ability of a system or component to function under stated conditions for a specified period of time[96]. The proximity of generative AI to patient well-being necessitates standards for reliability to minimize potential errors that could lead to accidents[43]. The generative AI models should have explicit and well-defined clinical use cases. A generative AI model designed for disease prediction needs to have a clear definition of the use situation and patient criteria. In addition, such generative AI models should be safe, secure, and effective throughout their life cycles. Generative AI models should be demonstrated to be at least as safe as human decision making alone and not cause undue harm. Existing generative AI models suffer from hallucination and output variations, undermining their ability to produce reliable outputs. These shortcomings can adversely affect the trust between physicians and generative AI systems. Adopting the DOD's principle for reliability can establish use cases for AI applications and monitor them during development and deployment to fix system failures and deterioration. Having a thorough evaluation and testing protocol against specific use cases will ensure the development of resilient and robust AI systems, and help minimize system failures as well as the time needed to respond to these errors.

Equity

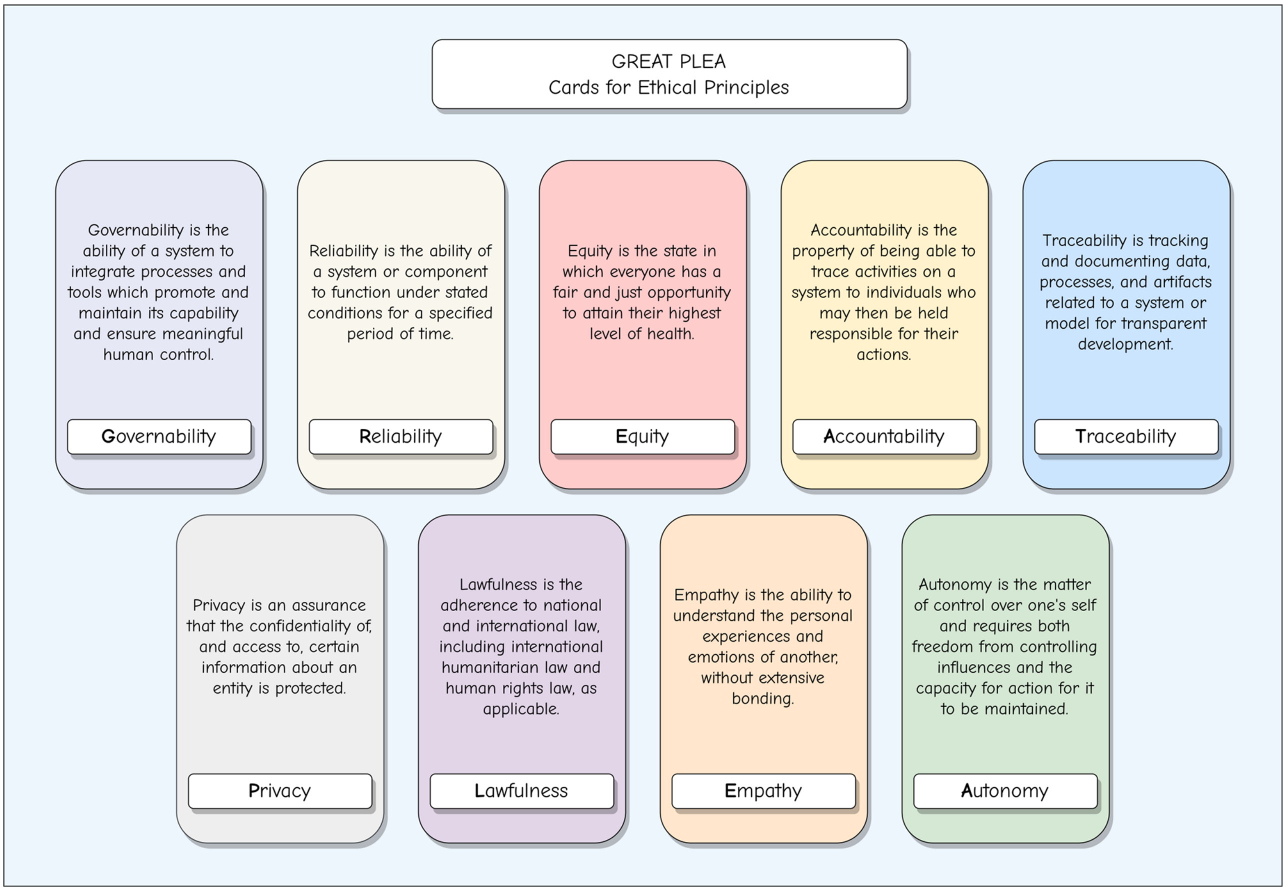
Equity is the state in which everyone has a fair and just opportunity to attain their highest level of health[97]. Due to the importance of health equity and the ramifications of algorithmic bias in healthcare, we call for adjustments to this principle. There already exists inequity in healthcare. The generative AI models, that naturally have elevated data bias risks due to their pre-training on massive datasets, should not exacerbate this inequity for marginalized, under-represented, socioeconomically disadvantaged, low education, or low health literacy groups[98], but rather incorporate their unique social situations into future AI models to insure equity. Generative AI must be developed with efforts to mitigate bias by accounting for existing health disparities. Without this consideration, generative AI systems could erroneously recommend treatments for different patients[99]. Expansion of the principle for equity must set standards for evaluation metrics of algorithmic fairness so that deployed AI systems will not reinforce healthcare disparity.

Accountability

Accountability is the property of being able to trace activities on a system to individuals who may then be held responsible for their actions[100]. To ensure accountability and human involvement with AI in healthcare, the principle of Responsibility and Accountability outlined by NATO[3] states that they will develop AI applications mindfully and integrate human responsibility to establish human accountability for actions taken by or with the application. A study of patient attitudes toward AI showed the importance of accountability in gaining patient trust when using AI in healthcare[101]. This assurance of accountability is crucial when a clinician is using generative AI to help treat a patient, as without proper measures for human accountability, the patient may feel that the clinician is not invested in the care they are delivering[102]. We can adopt this principle for the ethical use of generative AI in healthcare, and ensure that human involvement is maintained when more powerful generative AI systems, such as ChatGPT or generative AI-based clinical decision support systems, are used in patient care.

Traceability

Traceability is tracking and documenting data, processes, and artifacts related to a system or model for transparent development[103]. Addressing the issue of optimizing trust between healthcare professionals and the AI they interact with can be done by adopting the principle of traceability. This way, the personnel working with AI will understand its capabilities, developmental process, methodologies, data sources, and documentation. Furthermore, providing personnel with the understanding of an AI system capabilities and the processes behind its actions, will also improve system reproducibility, allowing for seamless deployment across healthcare systems. This is important for generative AI systems in healthcare because of their nature of being a black box system. This high-level understanding will help optimize trust, as operators will be aware of the capabilities and limitations of the AI systems they work with and know the appropriate settings for use[104]. With generative AI becoming more prevalent in healthcare, proper documentation is required to ensure all end users are properly educated on the capabilities and limitations of the systems they interact with. The generation process of generative AI models should be transparent. The references or facts should be provided together with answers and suggestions for clinicians and patients. Data sources used to train these models and the design procedures of these models should be transparent too. Furthermore, the implementation, deployment, and operation of these models need to be auditable, under the control of stakeholders in the healthcare setting.



GREAT PLEA
Cards for Ethical Principles

Governability is the ability of a system to integrate processes and tools which promote and maintain its capability and ensure meaningful human control.

Governability

Reliability is the ability of a system or component to function under stated conditions for a specified period of time.

Reliability

Equity is the state in which everyone has a fair and just opportunity to attain their highest level of health.

Equity

Accountability is the property of being able to trace activities on a system to individuals who may then be held responsible for their actions.

Accountability

Traceability is tracking and documenting data, processes, and artifacts related to a system or model for transparent development.

Traceability

Privacy is an assurance that the confidentiality of, and access to, certain information about an entity is protected.

Privacy

Lawfulness is the adherence to national and international law, including international humanitarian law and human rights law, as applicable.

Lawfulness

Empathy is the ability to understand the personal experiences and emotions of another, without extensive bonding.

Empathy

Autonomy is the matter of control over one's self and requires both freedom from controlling influences and the capacity for action for it to be maintained.

Autonomy

Fig. 4 GREAT PLEA cards for ethical principles. We propose the "GREAT PLEA" ethical principles for generative AI in healthcare, namely Governability, Reliability, Equity, Accountability, Traceability, Privacy, Lawfulness, Empathy, and Autonomy. The GREAT PLEA ethical principles demonstrate our great plea for the community to prioritize these ethical principles when implementing and utilizing generative AI in practical healthcare settings.

Privacy

Privacy is an assurance that the confidentiality of, and access to, certain information about an entity is protected[105]. Privacy is necessary in most military and medical applications of healthcare due to their confidential nature. Generative AI systems in healthcare must be HIPAA compliant for data disclosures, and secure to prevent breaches and developers should be advised how healthcare data should train systems for deployment. HIPAA compliance requires a regular risk assessment to determine how vulnerable patient data is ref. [106], thus a clinic utilizing generative AI systems in healthcare would have to determine if these systems are weak points in their technology network. For example, the utilization of generative AI models presents potential privacy breach risks, including prompt injection[107], where malicious actions could be conducted by overriding an original prompt, and jailbreak[108], where training data could be divulged by eliciting generated content. Furthermore, the capabilities of generative AI to process personal data and generate sensitive information make it crucial for these systems to be secure against data breaches and cyberattacks. Ensuring these systems are developed with data privacy and security in mind will assist in keeping protected patient information secure. Having these robust measures in place to maintain the privacy of the sensitive data collected and made by AI systems is crucial for the well-being of patients and for building trust with patients.

Lawfulness

Lawfulness is the adherence to national and international law, including international humanitarian law and human rights law, as applicable[109]. This can be adopted for the use of generative AI in healthcare. The laws that must be adhered to are not laws of conflict, but rather those related to healthcare. Different states in the U.S. may establish different laws for AI systems that must be heeded for deployment in those areas[110]. Generative AI systems in healthcare also face legal challenges surrounding safety and effectiveness, liability, data privacy, cybersecurity, and intellectual property law[92]. A legal foundation must be established for the liability of action taken and recommended by these systems, as well as considerations for how they interact with cybersecurity and data privacy requirements of healthcare providers. Generative AI for healthcare must be developed with these legal challenges in mind to protect patients, clinicians, and AI developers from any unintended consequences.

Empathy

Empathy is the ability to understand the personal experiences and emotions of another, without extensive bonding[111]. A principle for empathy is not directly referenced in any guidelines by the DOD or NATO. However, by emphasizing the need for human involvement in the treatment of patients, it is possible to create a framework for human involvement in generative AI applications to prevent gaps in accountability and ensure patients receive care that is empathetic and helpful[53]. There have been notable concerns about artificial empathy[112] of chatbots, such as ChatGPT, reinforcing the need for a principle defining empathy for generative AI in healthcare[113]. An empathetic relationship between provider and patient brings several benefits to both the patient and the clinic treating them, such as better patient outcomes, fewer disputes with healthcare providers, higher patient satisfaction, and higher reimbursement[111].

Autonomy

Autonomy is the matter of control over one's self and requires both freedom from controlling influences and the capacity for action for it to be maintained[114,115]. The more powerful AI systems become, the more concerns arise that humans do not control healthcare systems and care decisions[32]. Generative AI has seen staggering progress in the past several years, and hence, the protection of autonomy needs to be ensured when using generative AI in healthcare. Protecting human autonomy means that patients receive care according to their preferences and values and that clinicians can deliver treatment in the manner they want, without being encroached upon by the generative AI system. If autonomy in decision-making is not patient-focused, the potential for adverse events and poor clinical outcomes will surely follow[116]. By including provisions for protecting autonomy in using generative AI in healthcare, doctor-patient relations improve, and care quality is ultimately improved[117].

CONCLUSION

Generative AI has great potential to enhance and make high-quality healthcare more accessible to all, leading to a fundamental transformation in its delivery. Challenges posed by AI in healthcare often mirror those encountered in military. We propose the GREAT PLEA ethical principles, encompassing nine ethical principles, in the hope of addressing the ethical concerns of generative AI in healthcare, as well as the distinction between generative AI and "general" AI. This will be achieved by addressing the elevated risks mentioned previously in the paper. Generative AI necessitates guidelines that account for the risk of misinformation, ramifications of bias, and difficulty of using general evaluation metrics. Considering the widespread nature of generative AI and its risks, these ethical principles can protect patients and clinicians from unforeseen consequences. Following these principles, generative AI can be continuously evaluated for errors, bias, and other concerns that patients or caregivers may have about their relationship with AI in their field. The present moment urges us to embrace these principles, foster a closer collaboration between humans and technology, and effect a radical enhancement in the healthcare system.

These principles can be enforced through cooperation with lawmakers and the establishment of standards for developers and users, as well as a partnership with recognized governing bodies within the healthcare sector, such as the WHO or AMA.

We note that the enforcement of the proposed ethical principles, be it via evaluation approaches (e.g., Likert scale, prompting, or semantic similarity-based approaches for empathy[53,118,119]) or through other means, is out of the scope of this effort. As such, we acknowledge the lack of detailed enforcement procedures as the limitation of the work. At the same time, implementing AI metrics or enforcement methods for GREAT PLEA ethical principles can also be the potential future avenue for exploration.

Received: 18 May 2023; Accepted: 15 November 2023;
Published online: 02 December 2023

REFERENCES

1. Russell, S. Ai weapons: Russia's war in Ukraine shows why the world must enact a ban. *Nature* https://www.nature.com/articles/d41586-023-00511-5 (2023).
2. U.S. Department of Defense. Dod adopts ethical principles for artificial intelligence https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/ (2020).
3. The North Atlantic Treaty Organization. Summary of the NATO artificial intelligence strategy https://www.nato.int/cps/en/natohq/official_texts_187617.htm (2021).
4. Hicks, K. What the Pentagon thinks about artificial intelligence. *Politico* https://www.politico.com/news/magazine/2023/06/15/pentagon-artificial-intelligence-china-00101751.
5. Baker, A. et al. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Front Artif. Intell.* **3**, 543405 (2020).
6. Chan, S. & Siegel, E. L. Will machine learning end the viability of radiology as a thriving medical specialty? *Br. J. Radiol.* **92**, 20180416 (2019).
7. Meyer, J. et al. Impact of artificial intelligence on pathologists' decisions: an experiment. *J. Am. Med. Inform. Assoc.* **29**, 1688–1695 (2022).
8. Langlotz, C. P. Will artificial intelligence replace radiologists? *Radiol. Artif. Intell.* **1**, e190058 (2019).
9. Cacciamani, G. E. et al. Is artificial intelligence replacing our radiology stars? not yet! *Eur. Urol. Open Sci.* **48**, 14–16 (2023).
10. Yang, X. et al. A large language model for electronic health records. *npj Digit. Med.* **5**, 194 (2022).
11. Lin, W.-C., Chen, J. S., Chiang, M. F. & Hribar, M. R. Applications of artificial intelligence to electronic health record data in ophthalmology. *Transl. Vis. Sci. Technol.* **9**, 13–13 (2020).
12. Rosenthal, S., Barker, K. & Liang, Z. Leveraging medical literature for section prediction in electronic health records. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4864–4873 (Association for Computational Linguistics, Hong Kong, China, 2019).
13. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
14. Organization, T. W. H. Ethics and governance of artificial intelligence for health https://www.who.int/publications/i/item/9789240029200 (2021).
15. Dowling, M. & Lucey, B. Chatgpt for (finance) research: the Bananarama conjecture. *Finance Res. Lett.* **53**, 103662 (2023).
16. Lee, M., Liang, P. & Yang, Q. Coauthor: designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22 (Association for Computing Machinery, New York, NY, USA, 2022). https://doi.org/10.1145/3491102.3502030.
17. Thiergart, J., Huber, S. & Übellacker, T. Understanding emails and drafting responses—an approach using gpt-3 (2021). Preprint at https://arxiv.org/abs/2102.03062.
18. Ranade, P., Piplai, A., Mittal, S., Joshi, A. & Finin, T. Generating fake cyber threat intelligence using transformer-based models. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–9 (2021).
19. Liao, W. et al. Differentiate chatgpt-generated and human-written medical texts (2023). Preprint at https://arxiv.org/abs/2304.11567.
20. Chintagunta, B., Katariya, N., Amatriain, X. & Kannan, A. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, (eds Shivade, C. et al.) 66–76 (Association for Computational Linguistics, Online, 2021). https://aclanthology.org/2021.nlpmc-1.9.
21. Sun, Z. et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology* **307**, e231259 (2023).
22. Peng, Y., Rousseau, J. F., Shortliffe, E. H. & Weng, C. AI-generated text may have a role in evidence-based medicine. *Nat. Med.* (2023).
23. Gilbert, T. K., Brozek, M. W. & Brozek, A. Beyond bias and compliance: Towards individual agency and plurality of ethics in AI (2023). Preprint at https://arxiv.org/abs/2302.12149.
24. Birhane, A. et al. The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022).
25. OpenAI. Introducing chatgpt https://openai.com/blog/chatgpt (2022).
26. Hu, K. Chatgpt sets record for fastest-growing user base - analyst note. *Reuters* https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
27. OpenAI. Model index for researchers https://platform.openai.com/docs/model-index-for-researchers.
28. OpenAI. Gpt-4 technical report (2023). Preprint at https://arxiv.org/abs/2303.08774.
29. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. https://openai.com/research/language-unsupervised (2018).
30. Radford, A. et al. Language models are unsupervised multitask learners (2019).

31. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, Vol. 33 (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) 1877–1901 (Curran Associates, Inc., 2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
32. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30 (eds Guyon, I. et al.) (Curran Associates, Inc., 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685 (IEEE Computer Society, Los Alamitos, CA, USA, 2022). https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents (2022). Preprint at https://arxiv.org/abs/2204.06125.
35. Luo, C. Understanding diffusion models: A unified perspective (2022). Preprint at https://arxiv.org/abs/2208.11970.
36. Zhao, W. X. et al. A survey of large language models (2023). Preprint at https://arxiv.org/abs/2303.18223.
37. Liu, P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55** https://doi.org/10.1145/3560815 (2023).
38. Kather, J. N., Ghaffari Laleh, N., Foersch, S. & Truhn, D. Medical domain knowledge in domain-agnostic generative ai. *npj Digit. Med.* **5**, 90 (2022).
39. Zhang, C. et al. A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need? (2023). Preprint at https://arxiv.org/abs/2303.11717.
40. Zhang, C., Zhang, C., Zhang, M. & Kweon, I. S. Text-to-image diffusion models in generative ai: A survey (2023). Preprint at https://arxiv.org/abs/2303.07909.
41. Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models (2023). Preprint at https://arxiv.org/abs/2304.03738.
42. Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I. & Pauly, M. The self-perception and political biases of chatgpt (2023). Preprint at https://arxiv.org/abs/2304.07333.
43. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
44. Bang, Y. et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity (2023). Preprint at https://arxiv.org/abs/2302.04023.
45. Bian, N. et al. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models (2023). Preprint at https://arxiv.org/abs/2303.16421.
46. Chen, N. et al. Metrics for deep generative models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Vol. 84 of *Proceedings of Machine Learning Research*, (eds Storkey, A. & Perez-Cruz, F.) 1540–1550 (PMLR, 2018). https://proceedings.mlr.press/v84/chen18e.html.
47. Thoppilan, R. et al. Lamda: Language models for dialog applications (2022). Preprint at https://arxiv.org/abs/2201.08239.
48. Gloria, K., Rastogi, N. & DeGroff, S. Bias impact analysis of AI in consumer mobile health technologies: Legal, technical, and policy (2022). Preprint at https://arxiv.org/abs/2209.05440.
49. Peng, C. et al. A study of generative large language model for medical research and healthcare (2023). Preprint at https://arxiv.org/abs/2305.13523.
50. Wei, J. et al. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (eds Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K.) https://openreview.net/forum?id=_VjQlMeSB_J (2022).
51. Leiter, C. et al. Towards explainable evaluation metrics for natural language generation (2022). Preprint at https://arxiv.org/abs/2203.11131.
52. Priyanshu, A., Vijay, S., Kumar, A., Naidu, R. & Mireshghallah, F. Are chatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt-induced sanitization (2023). Preprint at https://arxiv.org/abs/2305.15008.
53. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* (2023).
54. Donovan - AI-powered decision-making for defense. *Scale* https://scale.com/donovan (2023).
55. Advanced targeting and lethality aided system (atlas). *CoVar* https://covar.com/case-study/atlas/ (2023).
56. Doctrinaire. *CoVar* https://covar.com/case-study/doctrinaire/ (2023).
57. Choudhury, A. & Asan, O. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med. Inform.* **8**, e18599 (2020).
58. Bahl, M. et al. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology* **286**, 170549 (2017).
59. Dalal, A. K. et al. Systems engineering and human factors support of a system of novel ehr-integrated tools to prevent harm in the hospital. *J. Am. Med. Inform. Assoc.* **26**, 553–560 (2019).
60. *Intercom for Healthcare* https://www.intercom.com/drlp/industry/healthcare.
61. *Prediction and Early Identification of Disease Through AI—Siemens Healthineers* https://www.siemens-healthineers.com/digital-health-solutions/artificial-intelligence-in-healthcare/ai-to-help-predict-disease.
62. Willemink, M. Ai for CT image reconstruction - a great opportunity. *AI Blog* https://ai.myesr.org/articles/ai-for-ct-image-reconstruction-a-great-opportunity/ (2019).
63. Bajgain, B., Lorenzetti, D., Lee, J. & Sauro, K. Determinants of implementing artificial intelligence-based clinical decision support tools in healthcare: a scoping review protocol. *BMJ Open* **13**, e068373 (2023).
64. David Lat, E. M. Advanced targeting and lethality automated system archives. *Breaking Defense* https://breakingdefense.com/tag/advanced-targeting-and-lethality-automated-system/.
65. Utegen, A. et al. Development and modeling of intelligent control system of cruise missile based on fuzzy logic. In *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, 1–6 (2021).
66. Bohr, A. & Memarzadeh, K. Chapter 2 - the rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare*, (eds Bohr, A. & Memarzadeh, K.) 25–60 (Academic Press, 2020).
67. Morgan, F. E. et al. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World* (RAND Corporation, Santa Monica, CA, 2020).
68. Introduction to the law of armed conflict (loac) https://www.genevacall.org/wp-content/uploads/dlm_uploads/2013/11/The-Law-of-Armed-Conflict.pdf.
69. Rule 1. The principle of distinction between civilians and combatants. *IHL* https://ihl-databases.icrc.org/en/customary-ihl/v1/rule1.
70. Docherty, B. Losing humanity. *Human Rights Watch* https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots (2012).
71. *Generative Artificial Intelligence and data privacy: A Primer - CRS Reports* https://crsreports.congress.gov/product/pdf/R/R47569.
72. Journal, H. Hipaa, healthcare data, and artificial intelligence. *HIPAA J.* https://www.hipaajournal.com/hipaa-healthcare-data-and-artificial-intelligence/ (2023).
73. Patel, V. L., Kannampallil, T. G. & Kaufman, D. R. *Cognitive informatics for biomedicine: human computer interaction in healthcare* (Springer, 2015).
74. II, W. N. P. Risks and remedies for artificial intelligence in health care. *Brookings* https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care/ (2022).
75. Lyons, J. B. & Stokes, C. K. Human-human reliance in the context of automation. *Hum. Factors* **54**, 112–121 (2012).
76. Asan, O., Bayrak, E. & Choudhury, A. Artificial intelligence and human trust in healthcare: Focus on clinicians (preprint) (2019).
77. Lewis, M., Sycara, K. & Walker, P. *The Role of Trust in Human–Robot Interaction*, 135–159 (Springer International Publishing, 2018).
78. Hawley, J. K. Looking back at 20 years of manprint on patriot: Observations and lessons (2007).
79. Parikh, R. B., Obermeyer, Z. & Navathe, A. S. Regulation of predictive analytics in medicine. *Science* **363**, 810–812 (2019).
80. Richardson, J. P. et al. Patient apprehensions about the use of artificial intelligence in healthcare. *npj Digit. Med.* **4**, 140 (2021).
81. Christian, R. Mind the gap the lack of accountability for killer robots. *Human Rights Watch* https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots (2015).
82. Habli, I., Lawton, T. & Porter, Z. Artificial intelligence in health care: accountability and safety. *Bull. World Health Organ.* **98**, 251–256 (2020).
83. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
84. N, O. et al. Addressing racial and ethnic inequities in data-driven health technologies 1–53 (2022).
85. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
86. Frisk, A. What is Project Maven? The Pentagon ai project Google employees want out of - -national. *Global News* (2018). https://globalnews.ca/news/4125382/google-pentagon-ai-project-maven/.
87. Shane, S. & Wakabayashi, D. The business of war': Google employees protest work for the Pentagon. *The New York Times* https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html (2018).
88. Our principles. *Google AI* https://ai.google/principles.

89. *Augmented intelligence in Health Care*1 - American Medical Association* https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf.
90. *Blueprint for trustworthy AI implementation guidance and assurance for healthcare* https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf.
91. Blueprint for an AI bill of rights - ostp. *The White House* https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (2023).
92. Naik, N. et al. Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility?*Front. Surg.* **9** (2022).
93. Pifer, R. "hurtling into the future": The potential and thorny ethics of generative ai in healthcare. *Healthcare Dive* https://www.healthcaredive.com/news/generative-AI-healthcare-gpt-potential/648104/ (2023).
94. Rosenberg, I., Shabtai, A., Elovici, Y. & Rokach, L. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Comput. Surv.* **54** https://doi.org/10.1145/3453158 (2021).
95. Sigfrids, A., Leikas, J., Salo-Pöntinen, H. & Koskimies, E. Human-centricity in AI governance: A systemic approach. *Front. Artif. Intell.* **6** https://www.frontiersin.org/articles/10.3389/frai.2023.976887 (2023).
96. Developing cyber-resilient systems: A systems security engineering approach https://doi.org/10.6028/NIST.SP.800-160v2r1.
97. *Centers for Disease Control and Prevention* https://www.cdc.gov/healthequity/whatis/index.html (2022).
98. Aquino, Y. S. J. et al. Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: a qualitative study of multidisciplinary expert perspectives. *J. Med. Ethics* (2023).
99. Hoffman, K. M., Trawalter, S., Axt, J. R. & Oliver, M. N. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc. Natl Acad. Sci.* **113**, 4296–4301 (2016).
100. Oldehoeft, A. E. Foundations of a security policy for use of the national research and educational network https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir4734.pdf.
101. Robertson, C. et al. Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. *PLOS Digital Health* https://doi.org/10.1371/journal.pdig.0000237.
102. Habli, I., Lawton, T. & Porter, Z. Artificial intelligence in health care: accountability and safety. *Bull. World Health Org.* **98**, 251 – 256 (2020).
103. Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. & Sicilia, M.-A. Traceability for trustworthy AI: a review of models and tools. *Big Data Cogn. Comput.* **5** https://www.mdpi.com/2504-2289/5/2/20 (2021).
104. Li, B. et al. Trustworthy ai: From principles to practices. *ACM Comput. Surv.* **55** https://doi.org/10.1145/3555803 (2023).
105. Barker, E., Smid, M., Branstad, D. & Chokhani, S. A framework for designing cryptographic key management systems https://csrc.nist.gov/publications/detail/sp/800-130/final.
106. (OCR), O. f. C. R. Guidance on risk analysis. *HHS.gov* https://www.hhs.gov/hipaa/for-professionals/security/guidance/guidance-risk-analysis/index.html (2021).
107. Perez, F. & Ribeiro, I. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop* https://openreview.net/forum?id=qiaRo_7Zmug (2022).
108. Liu, Y. et al. Jailbreaking chatgpt via prompt engineering: An empirical study (2023). Preprint at https://arxiv.org/abs/2305.13860.
109. Stanley-Lockman, Z. & Christie, E. H. An artificial intelligence strategy for nato https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.
110. Team, T. F. State of California endorses Asilomar ai principles. *Future Life Inst.* https://futureoflife.org/recent-news/state-of-california-endorses-asilomar-ai-principles/ (2022).
111. Moudatsou, M., Stavropoulou, A., Philalithis, A. & Koukouli, S. The role of empathy in health and social care professionals. *Healthcare* **8**, 26 (2020).
112. Zhu, Q. & Luo, J. Toward artificial empathy for human-centered design: A framework (2023). Preprint at https://arxiv.org/abs/2303.10583.
113. Asch, D. A. An interview with chatgpt about health care. *Catal. Non Issue Content* **4** (2023).
114. Holm, S. Principles of biomedical ethics, 5th edn. *J. Med. Eth.* **28**, 332–332 (2002).
115. *AMA Journal of Ethics* **18**, 12–17 (2016).
116. Applin, S. & Fischer, M. New technologies and mixed-use convergence: How humans and algorithms are adapting to each other (2016).
117. *Human Rights and Biomedicine* https://coe.int/en/web/bioethics/report-impact-of-ai-on-the-doctor-patient-relationship.
118. Svikhnushina, E. & Pu, P. Approximating online human evaluation of social chatbots with prompting. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, (eds Schlangen, D. et al.) 268–281 (Association for Computational Linguistics, 2023). https://aclanthology.org/2023.sigdial-1.25.
119. Raamkumar, A. S. & Yang, Y. Empathetic conversational systems: a review of current advances, gaps, and opportunities (2022). Preprint at https://arxiv.org/abs/2206.05017.

ACKNOWLEDGEMENTS

Y.W. would like to acknowledge support from the University of Pittsburgh Momentum Funds, Clinical and Translational Science Institute Exploring Existing Data Resources Pilot Awards, the School of Health and Rehabilitation Sciences Dean's Research and Development Award, and the National Institutes of Health through Grants UL1TR001857, U24TR004111, and R01LM014306. Y.P. would like to acknowledge support from the National Institutes of Health through Grants 4R00LM013001, R01LM014306, and the National Science Foundation through Grant 2145640. The sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

AUTHOR CONTRIBUTIONS

D.O. conceptualized, designed, and organized this study, analyzed the results, and wrote, reviewed, and revised the paper. J.H. analyzed the results, and wrote, reviewed, and revised the paper. R.K.P., J.C.P., G.L.L., and Y.P. wrote, reviewed, and revised the paper. Y.W. conceptualized, designed, and directed this study, wrote, reviewed, and revised the paper.

COMPETING INTERESTS

Y.W. consults for Pfizer Inc. and has ownership/equity interests in BonafideNLP, LLC. All other authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Yanshan Wang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.