Current Treatment Options in Psychiatry (2025) 12:5
https://doi.org/10.1007/s40501-024-00340-x

REVIEW

Generative Artificial Intelligence in Mental Healthcare: An Ethical Evaluation

Charlotte Blease[1,2] · Adam Rodman[3]

Accepted: 14 October 2024

Abstract

Purpose Since November 2022, generative artificial intelligence (AI) chatbots, such as ChatGPT, that are powered by large language models (LLM) have been the subject of growing attention in healthcare. Using biomedical ethical principles to frame our discussion, this review seeks to clarify the current ethical implications of these chatbots, and to identify the key empirical questions that should be pursued to inform ethical practice.

Recent findings In the past two years, research has been conducted into the capacity of generative AI chatbots to pass medical school examinations, evaluate complex diagnostic cases, solicit patient histories, interpret and summarize clinical documentation, and deliver empathic care. These studies demonstrate the scope and growing potential of this AI to assist with clinical tasks.

Summary Despite increasing recognition that generative AI can play a valuable role in assisting with clinical tasks, there has been limited, focused attention paid to the ethical consequences of these technologies for mental healthcare. Adopting a framework of biomedical ethics, this review sought to evaluate the ethics of generative AI tools in mental healthcare, and to motivate further research into the benefits and harms of these tools.

Keywords Large language models · Generative artificial intelligence · Ethics · Mental health · ChatGPT · Psychiatry · Psychotherapy

Introduction

Amid the growing global demand for improved access to mental health services, healthcare organizations, and patients, are increasingly turning to technological innovations to enhance care delivery and reduce costs [1]. While digital and artificial intelligence (AI) technologies for mental health have their roots in the 1960s [2], discussions of their role in the provision of mental health care have grown since the public release of OpenAI's ChatGPT in November 2022. While the concept of generative AI (GAI) – AI systems capable of creating human-like output – is not entirely new, recent advancements and the widespread availability of large language models (LLMs), such as OpenAI's GPT-4 and Google's Bard, suggest that these technologies could have important clinical applications. LLMs are a form of generative AI capable of analyzing and creating content by leveraging vast data troves including publicly accessible information on the internet. Unlike traditional search engines that return links in response to user queries, chatbots powered by these models can generate rapid outputs that 'remember' previous user exchanges and appear to mimic natural human conversations.

Emerging research suggests that psychiatrists and primary care physicians are adopting these tools to assist with clinical tasks [1, 3]. As early as June 2023, a Medical Economics survey conducted in the USA found that over 10% of clinicians had already started using chatbots like ChatGPT. Additionally, nearly 50% of respondents indicated plans to adopt these technologies in the future for tasks such as data entry, medical scheduling, or research [4]. By October 2023 a small survey (n = 138) conducted with psychiatrists affiliated to the American Psychiatric Association (APA) found that 44% of respondents had used ChatGPT 3.5 and 33% had used 4.0 *"to assist with*

✉ Charlotte Blease
charlotteblease@uu.se

1 Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

2 Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA

3 Beth Israel Deaconess Medical Center, Boston, MA, USA

answering clinical questions" [1]. Another study of 420 US medical students (response rate 50%), found that 40% had used ChatGPT [5]. Meanwhile, in February 2024 a major study of 1006 UK general practitioners found that 20% had used generative AI tools "*to assist with answering clinical questions"* with 16% specifically reporting the adoption of ChatGPT [3].

In the United States, health systems are rapidly integrating a variety of LLM-based tools into real clinical workflows. "Ambient listening," the use of an LLM-powered system that listens to patient-physician interactions and generates the first draft of a clinical note, has been implemented in many clinical settings across the country; early implementations of these systems have been optimistic, showing increases in patient and provider satisfaction without any serious safety concerns [6]. Similarly, response to patient messages, in which an LLM drafts that first draft of a message to a patient through an EHR portal, have been widely adopted, though with more mixed results [7]. EHR providers, most notably Epic, have committed to full integration of LLMs throughout almost every domain of clinical workflows, including documentation efficiency including clinical summarization, patient experience, population health management, and a variety of billing tools – over 60 implementations in development [8]. In the United States, these uses of LLMs are not currently subject to Food and Drug Administration oversight, and many LLM tools are exempt from Software as a Medical Device (SaMD) regulation [9].

In August 2024, in a KFF health tracking poll in the US, about one in six adults (17%) reported using AI chatbots at least once a month to find health information and advice, rising to one quarter of adults under age 30 (25%) [10]. However, there is scarce research into patients' experiences with, and opinions about, these tools in mental healthcare [11]. Furthermore, with limited evidence and lack of concrete guidance by medical organizations and regulators about the use of generative AI [12–14], mental health clinicians may be uncertain when it is appropriate to use them, what to advise patients, and what constitutes best practice.

In this review, our aim is to go beyond current commentaries on generative AI in mental healthcare [15, 16] to evaluate and motivate further research into the benefits and harms of these tools, in this paper we use a framework of biomedical ethics [17]. We also identify key empirical questions in this domain that warrant further study (see Table 1).

Discussion

Respect for patient autonomy

Clinicians are obliged to be open and honest with patients and to respect their autonomy to make informed choices about their care. Respect for autonomy requires offering patients relevant, accessible, and timely information about their health and treatment options, in contexts that are free from coercion, so that individuals can exercise their right to healthcare determination.

Although respect for autonomy is a fundamental principle of medical ethics, research indicates that patients often misunderstand or forget substantial portions of the information conveyed to them during medical visits [18]. Clinicians, as experts in their field, often overestimate patients' understanding of specialized or technical terms. They frequently fail to adjust their language to match a patient's level of comprehension [19], a phenomenon known as "the curse of expertise" [20]. This creates a knowledge gap, making it difficult for patients to fully grasp the information including those housed in their electronic health records. Additionally, clinical records have historically been designed to serve as an aide memoir for clinicians or to communicate detailed medical information between other healthcare professionals, rather than to offer easily understandable information to patients.

A major strength of generative AI is its capacity to rapidly generate summaries of complex data and content and translate such information into requested literacy levels and tone. Such capacities may not only render clinicians' administrative tasks more efficient, in the era of patient online record access they may assist clinicians in writing clinical documentation that patients can better understand [21]. When it comes to informed consent, ethicists have argued that LLMs could, in principle, improve patients' access to the relevant procedural information, therefore enhancing informed decision-making [22]. For this to occur, however, generative AI will need to furnish patients with information that is at least more accurate, accessible, and trustworthy than that proffered by clinicians in traditional consent scenarios [22, 23].

How reliable and trustworthy are generative AI chatbots? There is a well-documented tendency for LLMs to make up false information, referred to as 'hallucinations' [24], some of which may be subtly incorrect. Important to clarify is that many widely available chatbots such as ChatGPT are not specifically trained on medical data, and medical-grade models, such as Google's PALMMed2, exhibit higher medical fidelity [25]. Computational techniques such as retrieval-augmented generation (RAG) have also been shown to meaningfully reduce hallucination rate [26]. Nonetheless, even these models can still be prone to errors even while there is evidence that they are improving [27]. Moreover, due to ChatGPT's commercial availability and its widespread adoption as the most commonly used LLM chatbot with early studies showing that physicians are already utilizing it [3]. The rapid responses, and authoritative tone of conversational responses generated by LLM-powered chatbots could make both clinicians and patients more susceptible to misinformation, potentially undermining the quality of

Table 1 Ethical issues that may be informed by empirical research

Ethical Principle	Suggested Empirical Research Questions*	
	Following access to generative AI…	
	Patients' experiences**	Clinician experiences***
Respect for Patient Autonomy	Do patients better understand their healthcare? Do these tools improve patient understanding about medications? Do patients feel more empowered?	Do clinicians communicate information more understandably in narrative notes? Do clinicians preserve the clinical detail in their documentation?
Beneficence		
Empathic care & the therapeutic alliance *Diagnostic accuracy*	Do patients perceive responses written wholly by generative AI to be empathic? Do patients perceive responses cowritten between clinicians and generative AI to be empathic? Does the use of generative AI strengthen the therapeutic alliance with clinicians? Do these tools improve wellbeing? Do patients with mental health conditions use these tools for diagnostic purposes before visiting a clinician? How do patients with psychiatric disorders, such as schizophrenia, perceive the impact of generative AI-assisted diagnostic tools on the quality of care they receive, particularly in addressing both mental health and physical health conditions? Do patients use these tools as second opinions, helping to improved diagnostic accuracy?	Does the use of generative AI reduce compassion fatigue? Do clinicians find these tools useful for brainstorming diagnostics? Do these tools improve diagnostic accuracy rates, and reduce diagnostic overshadowing in mental healthcare?
Nonmaleficence		
Medical errors *Misinformation that leads to harm*	Do patients perceive errors in generative AI outputs? Does use increase patient anxiety? Does use increase self-harm episodes?	Do clinicians perceive errors in generative AI outputs? Does use of generative AI reduce medical error rates? Do clinicians rely on misinformation that later leads to patient harm?
Justice		
Unfair treatment *Access to care*	Do patients perceive stigmatized language in responses? Do patients feel offended by what they read? How do patients from different demographics perceive these tools? Do these tools improve access to health information/clinicians? Are some patient demographics more inclined to use these tools?	Do clinicians use these tools to reduce risks of including stigmatized or offensive language in documentation? How does these tools affect the fairness of treatment decisions across different patient demographics?
Privacy & Confidentiality	Do patients feel more worried about privacy? What do patients understand about how their health information will be used?	What do clinicians understand about privacy and confidentiality with respect to patient information when using generative AI tools?

* Empirical research should differentiate between different generative AI tools including medical grade tools, those that are specifically designed to be compliant with health privacy standards, and more commercially platforms. We envisage that a variety of methodologies should be applied to tackle these research questions including mixed methods survey research, natural language processing, and other techniques aimed at understanding objective measures of changes. We strongly recommend that, where feasible, randomized controlled trials are used to comparing human clinicians versus generative AI versus human clinicians + generative AI

**We recommend that research investigates the experiences and perceptions of patients with different mental health diagnoses

***We recommend research is conducted in different mental health settings with different healthcare professionals, including psychiatrists, clinical psychologists, psychotherapists, and mental health nurses disclosures, thereby compromising patient autonomy (see also *Privacy and confidentiality*).

Currently, only a handful of studies have investigated the patient communication abilities of content produced by generative AI with mixed findings. Tu et al. showed that a conversation AI system, AMIE, could take a clinical history better than human clinicians, though mental health care was excluded [28]. Pradhan et al. investigated the use of ChatGPT to write educational materials for cirrhosis and concluded that responses offered comparable readability, grade level, understandability, and accuracy to human-derived materials [29]. A study by Kharko et al. of primary care notes found that medical fidelity ratings varied, with ChatGPT 4.0 superior to version 3.5; ChatGPT required higher reading grades than the original primary care notes despite prompts requesting that the chatbot render the notes more accessible [30]. In contrast, another study reported that ChatGPT offered potential as a reliable source of psychoeducation, particularly among patients with very limited access to mental health resources [31].

Despite encouraging preliminary investigations, research into patient perspectives particularly in mental healthcare is limited [11]; whether access to generative AI improves patients' understanding and awareness about mental health conditions, including treatment options, is not yet fully understood. We recommend that future empirical research explore patients' sense of empowerment with generative AI, including how these tools influence quality of understanding following access. Experimental studies are also needed to assess the effectiveness of these tools in supporting clinical documentation, as well as assisting in taking patient histories, particularly in evaluating their responses to various prompts.

Beneficence

Empathic care and the therapeutic alliance

Sustaining consistently high levels of empathy within care delivery can be challenging, especially in mental health contexts where clinicians are particularly vulnerable to burnout ad compassion fatigue [32, 33]. Preliminary research suggests that LLMs might assist the delivery of empathy [34]. For example, a blinded study with clinician raters which compared responses from physicians and ChatGPT to 195 real-world health questions posted on Reddit's AskDocs reported ChatGPT's were rated as nearly 10 times more empathetic than the physicians' responses [35]. Other studies indicate that LLM-powered chatbots could help mental health professionals or peer supporters consistently provide high-quality support in patient interactions, especially among those dealing with compassion fatigue. For example, a randomized controlled trial involving responses on TalkLife, a peer-support social media platform for mental health, found that replies written in collaboration with a chatbot called 'HAILEY' (Human-AI coLlaboration approach for EmpathY) were more likely to be perceived as empathetic compared to human-only responses [36]. Peer supporters who reported struggling with empathy were significantly more likely to provide empathetic responses in the AI-assisted scenario.

Another study of ChatGPT explored its ability to translate fictional primary care notes, including a clinical note on major depressive disorder for a suicidal teenager, into more patient-friendly language [30]. Using the prompt, "Write an understandable and empathetic clinical note for the patient described in this record" the study found that the chatbot-generated notes contained significantly more markers of compassion, cognitive empathy, and pro-social cues compared to fictionalized notes written by a physician which exhibited negligible signatures of empathy.

Although chatbots have been found to demonstrate significantly more cues of empathy, particularly in written communication, it remains uncertain whether patients perceive these responses as genuinely empathic: blind assessments leave the actual patient perspective underexplored. Conceivably, if patients are not informed that a chatbot, rather than a human, is responding to their questions, it could undermine trust and the strength of the therapeutic alliance. For instance, in January 2023, the mental health platform Koko issued a public apology after using ChatGPT to generate emotional responses, misleading users into believing the replies were written by humans [37]. Further research is required to investigate patients' perspectives on clinical documentation produced or co-created by generative AI. For instance, future studies could examine how patients interpret "empathy" when it is conveyed by AI chatbots. We strongly recommend that empathy is carefully deconstructed as a concept in empirical research [38], and where feasible, validated measures examining the strength of the therapeutic alliance are used. In addition, studies could usefully investigate how this AI influences clinician burnout and compassion fatigue, and how patients perceive clinicians who adopt these tools in their communications.

Diagnostic accuracy

Current research shows that negative attitudes toward patients with psychiatric disorders [39], or the misattribution of physical symptoms to mental health conditions, can lead to errors in care [40] (see also: *Unfair treatment*, below). For example, patients with both serious mental illness and diabetes who visit emergency departments are less likely to be admitted for diabetic complications [41]. Additionally, hospitalized patients with schizophrenia face significantly higher risks of certain complications compared to those without the condition [42]. One promising use of generative AI in mental health care is its ability to assist clinicians with hypothesis generation, potentially overcoming risks associated with diagnostic overshadowing. Early research has shown that GPT-4 can produce accurate lists of differential diagnoses, even in complex cases [43, 44], which suggests its potential for supporting brainstorming in both diagnostic and treatment planning in mental health contexts. However, whether – in practice – generative AI augments or encumbers clinicians in making mental health diagnoses is unknown. We recommend that future research, including experimental studies, randomized controlled trials, retrospective case reviews, and patient surveys, explore the influence of generative AI in medical decisions including clinical outcome measures.

Nonmaleficence

Medical errors

A goal for AI in healthcare lies in harnessing its potential for personalized psychiatry, aiming to guide treatments that lead to better patient outcomes. As noted, however, the tendency for hallucinations may cause challenges, and a recent study in oncology reveals a growing concern: when ChatGPT 3.5 was prompted to provide cancer treatment suggestions, it frequently blended accurate information with incorrect recommendations, making it challenging even for experts to identify mistakes [45]. As noted, widely accessible commercial models like GPT-4 are not intended for medical purposes and although medical grade AI outperforms these chatbots, risks of medical error may still arise [28, 46].

Studies show some medically trained bots and humans perform at similar levels [28, 46]. For example, human experts found that 0.8 percent of Med-PaLM's answers included inappropriate biases, compared to 1.4 percent of clinicians' responses [46]. However, the issue of "hallucinations" was evident: clinicians provided incorrect information 1.4 percent of the time, while Med-PaLM did so in 18.7 percent of responses, and Flan-PaLM in 16.1 percent. Similar rates of incorrect information have been seen in generalist chatbots, such as GPT-4 [44].

We strongly recommend further empirical research is aimed at exploring the error rate of both medical grade generative AI tools, and more commercially available tools that may be more likely to be adopted by patients. When it comes to errors, the temporal consistency of tools, the types of errors they are liable to make compared with humans (an error ontology), and the corresponding error rates, should be explored.

Misinformation that leads to harm

Although the extent to which patients are adopting generative AI to seek health information is unclear, it is conceivable that these tools may sometimes offer inappropriate 'advice' that could risk leading to increased anxiety or even episodes of self-harm. For example, the possible adverse effects inflicted on the eating disorder community by the public release and swift withdrawal of the Tessa chatbot, within just one week, underscores the need for more comprehensive and reliable evidence than what has been gathered so far [47].

We are aware that no research has systematically explored the question of harm from generative AI with precision. Patient surveys could usefully explore the potential for negative experiences following generative AI usage. Tracking trends in the responses of patients to these tools, according to different mental health diagnoses or conditions, will be imperative. Automated scalable oversight systems, in which as AI model provides limited human-like oversight of another AI system, will likely be necessary given the extent of information to review.

Justice

Unfair treatment

The nature of the datasets used to train AI models is crucial, as any biases present in these datasets, or among the individuals involved in labeling or training the AI, are fundamentally embedded into the responses generated. Additionally, many widely accessible LLM-powered chatbots are not solely trained on medical literature and often handle information from the internet without differentiating its quality. Compounding this, the underrepresentation of women, racial and ethnic minorities, and seniors in research can lead to existing disparities in published medical texts [48–52]. Some studies suggest that these algorithmic biases could exacerbate discrimination in clinical practice [53, 54]. However, these important studies lack a human baseline, and more research is needed to examine whether algorithmic recommendations lead to worse bias than human-mediated care (see *Medical Errors*). Conceivably, for example, there may be circumstances when face-to-face discrimination among patients with mental health conditions [39] may be averted because interactions with chatbots may avoid the need for in-person or telemedicine encounters. In addition, LLMs can also help monitor the extent of discrimination in care by evaluating linguistic markers of biases [55], such as those found in clinical documentation [56].

Empirical research should explore the uptake, outcomes, and experiences of patients with generative AI across different mental health conditions, sex, age, and demographic groups. We recommend that attention is paid to risks of bias and treatment discrimination, among patients, and clinicians who adopt these tools.

Access to care

Justice in healthcare can also refer to ensuring fair and equal access to medical services, regardless of socioeconomic status, location, or background. Because of the ease of access of generative AI tools, particularly commercially available chatbots compared with accessing clinicians, there may be novel opportunities for patients to avail themselves of medical information. In the survey of APA-affiliated psychiatrists, 75% (n = 104) believed that patients would first consult these tools before first seeing a doctor [1]. Such access may be especially important among underserved populations and those who may lack health insurance coverage or those who are confronted by additional, cumbersome barriers to visiting clinicians. Again, patient experiences and preferences, including whether access to health information is elevated by these tools, is not yet understood. Without additional, multilevel efforts, generative AI could contribute to, and potentially exacerbate, the "digital divide" – whereby people who lack the ability or means to use internet-enabled tools.

Empirical research efforts should be focused on evaluating whether generative AI dilates access to high quality medical information, including preventative care, and examine the demographics of patients who are early adopters.

Privacy and confidentiality

Privacy is a key consideration in the patient-clinician relationship, and providers must communicate how confidentiality will be protected, including when clinicians are legally required to share information with third parties. Use of commercial LLM-powered chatbots such as ChatGPT, that do not adhere to medical privacy standards, can pose risks to patient privacy [23, 57]. These risks will vary across different legal systems due to the differing standards associated with processing identifiable, sensitive health information obtained from the internet [58, 59]. Nonetheless, due to the user-friendly interfaces, conversational nature, and the tendency of users to anthropomorphize these tools consent processes may be compromised leading to the inclination of users to share sensitive information [23]. Added privacy vulnerabilities arise because of the need to seek readily accessible assistance, advice, or information among those with mental health conditions who may fear stigmatization from clinicians, or who may otherwise be unable to easily access mental health care. In the survey conducted with APA-affiliated psychiatrists, more than half (57%, n = 79) anticipated that patients would worry more about privacy with these tools [1]. Once again, surveys of patients' experiences and opinions are lacking. We suggest that future quantitative and qualitative survey research should explore, more deeply, clinicians and patients' understanding and opinions about privacy and confidentiality with respect to generative AI tools.

Conclusion and recommendations

Generative AI is here to stay and it is poised for, and is already in widespread use, in drafting and co-authoring clinical documentation, and assisting clinicians with history-taking, diagnostics, empathy, and other clinical tasks [1, 3, 21, 28, 60]. However, we stress that current generative AI tools still present evidence-based risks, including inaccuracies, inconsistencies, hallucinations, and the potential to introduce harmful biases into clinical decision-making. Patients with mental health conditions are among the most vulnerable patients. They may also be turning to generative AI to supplement, or even replace health information and the emotional support traditionally derived from mental health clinicians and other health professionals.

Among our empirical research recommendations (see Table 1) we urge that investigators should differentiate between different generative AI tools including medical grade tools, those that are specifically designed to be compliant with health privacy standards, and more commercial platforms. Researchers should also avoid "in silico" evaluations – studying these tools with standardized text benchmarks – and instead broaden their studies to include how using these tools changes human behaviour, human–computer interaction. While the landscape appears to be changing at a faster rate than scientific evaluation, we also strongly recommend the use of randomized controlled trials to compare the effectiveness of human clinicians, generative AI, and the combination of both. When RCTs cannot feasibly be performed, we recommend robust evaluation within a quality-improvement paradigm. It is crucial to harness empirical evidence to inform ethical concerns about how best to weigh up the benefits and reduce harm that these tools can confer on patient care. In this paper, we have aimed to chart a clear path forward in evaluating ethical progress.

Key references

- Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ

Health & Care Informatics BMJ Publishing Group Ltd; 2024;31(1). Available from: https://informatics.bmj.com/content/31/1/e101102?fbclid=IwZXh0bgNhZW0CMTAAAR1NsNCk0SJ4upehd5KRl7VdDlCMkaXP090M2vh7Cln86x5bMmy3Nk_0f5g_aem_aq-vf4NA0Z5z4dxcCrLHFw [accessed Oct 7, 2024].

The largest survey to directly explore doctors' opinions and use of generative AI in clinical tasks.

- Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S, Singhal K, Cheng Y, Hou L, Webson A, Kulkarni K, Mahdavi SS, Semturs C, Gottweis J, Barral J, Chou K, Corrado GS, Matias Y, Karthikesalingam A, Natarajan V. Towards Conversational Diagnostic AI. arXiv; 2024. Available from: http://arxiv.org/abs/2401.05654 [accessed Jan 18, 2024].

  Notable study on the use of medical grade generative AI as a conversational agent in health care.

- Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine American Medical Association; 2023;183(6):589–596.

  Landmark randomized controlled trial comparing perceptions of empathy between generative AI and human doctors.

- Ferryman K, Mackintosh M, Ghassemi M. Considering Biased Data as Informative Artifacts in AI-Assisted Health Care. Drazen JM, editor. N Engl J Med 2023 Aug 31;389(9):833–838. doi: https://doi.org/10.1056/NEJMra2214964.

  The potential benefits of AI when it comes to identifying and overcoming discrimination.

Author contributions CB wrote the first draft and prepared Table 1 and AR contributed revisions. All authors reviewed and approved the final manuscript.

Funding Open access funding provided by Uppsala University. This work was supported by a FORTE grant AI in Healthcare Unleashed: Responsible and Ethical Implementation of Large Language Model Chatbots in Clinical Workflows and Patient Care [#2024–00039].

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. Psychiatry Res. 2024;333:115724.
2. Khawaja Z, Bélisle-Pipon JC. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. Front Digit Health. 2023;5:1278186.
3. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform. 2024;31(1). https://informatics.bmj.com/content/31/1/e101102?fbclid=IwZXh0bgNhZW0CMTAAAR1NsNCk0SJ4upehd5KRl7VdDlCMkaXP090M2vh7Cln86x5bMmy3Nk_0f5g_aem_aq-vf4NA0Z5z4dxcCrLHFw Accessed 7 Oct 2024.
4. Shryock T. AI Special Report: What patients and doctors really think about AI in health care. Med Econ. 2023. https://www.medicaleconomics.com/view/ai-special-report-what-patients-and-doctors-really-think-about-ai-in-health-care [accessed Aug 22, 2023].
5. Hosseini M, Gao CA, Liebovitz DM, Carvalho AM, Ahmad FS, Luo Y, MacDonald N, Holmes KL, Kho A. An exploratory survey about using ChatGPT in education, healthcare, and research. Plos one. 2023;18(10):e0292216.
6. Haberle T, Cleveland C, Snow GL, Barber C, Stookey N, Thornock C, Younger L, Mullahkhel B, Ize-Ludlow D. The impact of nuance DAX ambient listening AI documentation: a cohort study. J Am Med Inform Assoc. 2024;31(4):975–9.
7. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. JAMIA Open. 2024;7(2):ooae028.
8. EPIC. Artificial Intelligence. EPIC. 2024. https://www.epic.com/software/ai/ Accessed 7 Oct 2024.
9. Goodman KE, Paul HY, Morgan DJ. AI-Generated Clinical Summaries Require More Than Accuracy. JAMA. 2024. https://jamanetwork.com/journals/jama/article-abstract/2814609 Accessed 9 Apr 2024.
10. Presiado M, Montero A, Lopez L, Hamel L. KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information. KFF. 2024. https://www.kff.org/health-misinformation-and-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/ Accessed 13 Sep 2024.
11. Melo A, Silva I, Lopes J. Chatgpt: A pilot study on a promising tool for mental health support in psychiatric inpatient care. Int J Psychiatr Train. 2024. https://ijpt.scholasticahq.com/article/92367-chatgpt-a-pilot-study-on-a-promising-tool-for-mental-health- Accessed 12 Sep 2024.
12. American Psychiatric Association. The Basics of Augmented Intelligence: Some Factors Psychiatrists Need to Know Now. American Psychiatric Association. 2023. https://www.psychiatry.

org/News-room/APA-Blogs/The-Basics-of-Augmented-Intelligence Accessed 13 Aug 2023.
13. American Medical Association. AMA Augmented Intelligence Research Physician sentiments around the use of AI in heath care: motivations, opportunities, risks, and use cases. 2023. https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf Accessed 6 Aug 2024.
14. NHS England. Artificial Intelligence. NHS England. 2023. https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence/#:~:text=Guidance%20for%20healthcare%20workers&text=If%20you%20are%20using%20AI,via%20your%20clinical%20management%20route Accessed 15 Apr 2024.
15. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health. 2023;26(1):e300884. https://doi.org/10.1136/bmjment-2023-300884.
16 Torous J, Blease C. Generative artificial intelligence in mental health care: potential benefits and current challenges. World Psychiatry. 2024;23(1):1.
17. Beauchamp TL, Childress JF. Principles of biomedical ethics. USA: Oxford University Press; 2001.
18 McCarthy DM, Waite KR, Curtis LM, Engel KG, Baker DW, Wolf MS. What did the doctor say? Health literacy and recall of medical instructions. Medical Care. 2012;50(4):277.
19 Castro CM, Wilson C, Wang F, Schillinger D. Babel babble: physicians' use of unclarified medical jargon with patients. Am J Health Behav. 2007;31(1):S85–95.
20 Fisher M, Keil FC. The curse of expertise: When more knowledge leads to miscalibrated explanatory insight. Cogn Sci. 2016;40(5):1251–69.
21 Blease C, Torous J, McMillan B, Hägglund M, Mandl KD. Generative language models and open notes: exploring the promise and limitations. JMIR Med Educ. 2024;10:e51183.
22 Allen JW, Earp BD, Koplin J, Wilkinson D. Consent-GPT: is it ethical to delegate procedural consent to conversational AI? J Med Ethics. 2024;50(2):77–83.
23. Blease C. Open AI meets open notes: surveillance capitalism, patient privacy and online record access. J Med Ethics. 2024;50(2):84–9.
24. Goddard J. Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers. Am J Med. 2023. https://www.amjmed.com/article/S0002-9343(23)00401-1/abstract Accessed 6 Nov 2023.
25. Chen A, Chen DO. Accuracy of Chatbots in Citing Journal Articles. JAMA Netw Open. 2023;6(8):e2327647.
26. Kang H, Ni J, Yao H. Ever: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification. arXiv; 2024. http://arxiv.org/abs/2311.09114 Accessed 7 Oct 2024.
27 Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF. Evaluating large language models on medical evidence summarization. NPJ Digit Med. 2023;6(1):158.
28. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, Azizi S, Singhal K, Cheng Y, Hou L, Webson A, Kulkarni K, Mahdavi SS, Semturs C, Gottweis J, Barral J, Chou K, Corrado GS, Matias Y, Karthikesalingam A, Natarajan V. Towards Conversational Diagnostic AI. arXiv; 2024. http://arxiv.org/abs/2401.05654 Accessed 18 Jan 2024.
29 Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. Hepatol Commun. 2024;8(3):e0367.
30. Kharko A, McMillan B, Hagström J, Davidge G, Hagglund M, Blease C. Generative artificial intelligence writing open notes: A mixed methods assessment of the functionality of GPT 3.5 and GPT 4.0. Digital Health. 2024. https://doi.org/10.1177/20552076241291384.
31. Maurya RK, Montesinos S, Bogomaz M, DeDiego AC. Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. Couns and Psychother Res 2024;capr.12759. https://doi.org/10.1002/capr.12759.
32. Imo UO. Burnout and psychiatric morbidity among doctors in the UK: a systematic literature review of prevalence and associated factors. BJPsych Bull. 2017;41(4):197–204.
33. Bloom P. Against empathy: The case for rational compassion. New York City: Random House; 2017.
34. Inzlicht M, Cameron CD, D'Cruz J, Bloom P. In praise of empathic AI. Trends Cogn Sci. 2023. https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(23)00289-9 Accessed 31 May 2024.
35. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96.
36 Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell. 2023;5(1):46–57.
37. Ingram D. A mental health tech company ran an AI experiment on real users. Nothing's stopping apps from conducting more. NBC News. 2023. Available from: https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110 Accessed 13 Aug 2023.
38. Gerger H, Munder T, Kreuzer N, Locher C, Blease C. Lay perspectives on empathy in patient-physician communication: An online experimental study. Health Commun. 2024;39(6):1246–55.
39. Teachman BA, Wilson JG, Komarovskaya I. Implicit and explicit stigma of mental illness in diagnosed and healthy samples. J Soc Clin Psychol. 2006;25(1):75–95.
40. Shefer G, Henderson C, Howard LM, Murray J, Thornicroft G. Diagnostic overshadowing and other challenges involved in the diagnostic process of patients with mental illness who present in emergency departments with physical symptoms–a qualitative study. PLoS ONE. 2014;9(11):e111682.
41. Sullivan G, Han X, Moore S, Kotrla K. Disparities in hospitalization for diabetes among persons with and without co-occurring mental disorders. Psychiatr Serv. 2006;57(8):1126–31.
42. Daumit GL, Pronovost PJ, Anthony CB, Guallar E, Steinwachs DM, Ford DE. Adverse events during medical and surgical hospitalizations for persons with schizophrenia. Arch Gen Psychiatry. 2006;63(3):267–72.
43. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78–80.
44. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour R-E, Rodman A. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Intern Med. 2024. https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2817046 Accessed 30 Apr 2024.
45 Chen S, Kann BH, Foote MB, Aerts HJ, Savova GK, Mak RH, Bitterman DS. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol. 2023;9(10):1459–62.
46. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
47. Sharp G, Torous J, West ML. Ethical challenges in AI approaches to eating disorders. J Med Internet Res. 2023: e50696. https://www.jmir.org/2023/1/e50696/ Accessed 11 Sep 2024.
48 Dijkstra AF, Verdonk P, Lagro-Janssen AL. Gender bias in medical textbooks: examples from coronary heart disease, depression, alcohol abuse and pharmacology. Med Educ. 2008;42(10):1021–8.

49. Duma N, Vera Aguilera J, Paludo J, Haddox CL, Gonzalez Velez M, Wang Y, Leventakos K, Hubbard JM, Mansfield AS, Go RS. Representation of minorities and women in oncology clinical trials: review of the past 14 years. J Oncol Pract. 2018;14(1):e1–10.
50 Geller SE, Koch A, Pellettieri B, Carnes M. Inclusion, analysis, and reporting of sex and race/ethnicity in clinical trials: have we made progress? J Womens Health. 2011;20(3):315–20.
51 Watts G. Why the exclusion of older people from clinical research must stop. Bmj. 2012;344:e3445.
52. Bourgeois FT, Olson KL, Tse T, Ioannidis JP, Mandl KD. Prevalence and characteristics of interventional trials conducted exclusively in elderly persons: a cross-sectional analysis of registered clinical trials. PloS One. 2016;11(5):e0155948.
53. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdulnour RE, Butte AJ. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet Digital Health. 2024;6(1):e12–22.
54. Deb B, Rodman A. Racial Differences in Pain Assessment and False Beliefs About Race in AI Models. JAMA Netw Open. 2024;7(10):e2437977. https://doi.org/10.1001/jamanetworkopen.2024.37977.
55 Ferryman K, Mackintosh M, Ghassemi M. Considering Biased Data as Informative Artifacts in AI-Assisted Health Care. N Engl J Med. 2023;389(9):833–8. https://doi.org/10.1056/NEJMra2214964.
56 Himmelstein G, Bates D, Zhou L. Examination of Stigmatizing Language in the Electronic Health Record. JAMA Netw Open. 2022;5(1):e2144967.
57. Marks M, Haupt CE. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. JAMA. 2023;330(4):309–310. https://doi.org/10.1001/jama.2023.9458.
58. European Council of the European Union. Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world. 2023. https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ Accessed 11 Dec 2023.
59. Biden JR. Executive order on safe, secure, and trustworthy artificial intelligence. Federal Register. 2024;89(234):12345–50. https://www.federalregister.gov.
60. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, Singhal K, Sharma Y, Azizi S, Kulkarni K, Hou L, Cheng Y, Liu Y, Mahdavi SS, Prakash S, Pathak A, Semturs C, Patel S, Webster DR, Dominowska E, Gottweis J, Barral J, Chou K, Corrado GS, Matias Y, Sunshine J, Karthikesalingam A, Natarajan V. Towards Accurate Differential Diagnosis with Large Language Models. arXiv; 2023. http://arxiv.org/abs/2312.00164 Accessed 5 Apr 2024.

Artificial Intelligence (AI): Tools and Resources

Glossary of Artificial Intelligence Terms

Glossary provided by CIRCLS - The Center for Integrative Research in Computing and Learning Sciences.

This glossary was written for educators to reference when learning about and using artificial intelligence (AI).

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Adaptive Learning: Subject or course material is adjusted based on the performance of the learner. The difficulty of material, the pacing, sequence, type of help given, or other features can be adapted based on the learner's prior responses.

Algorithm: Algorithms are the "brains" of an AI system and what determines decisions. In other words, algorithms are the rules for what actions the AI system takes. Machine learning algorithms can discover their own rules (see machine learning for more) or be rule-based where human programmers give the rules.

Artificial General Intelligence (AGI): Artificial general intelligence has not yet been realized and would be when an A system can learn, understand, and solve any problem that a human can.

Artificial Intelligence (AI): AI is a branch of computer science. AI systems use hardware, algorithms, and data to create "intelligence" to do things like make decisions, discover patterns, and perform some sort of action. AI is a general term and there are more specific terms used in the field of AI. AI systems can be built in different ways, two of the primary ways are: (1) through the use of rules provided by a human (rule-based systems); or (2) with machine learning algorithms. Many newer AI systems use machine learning (see definition of machine learning below).

Artificial Narrow Intelligence (ANI): AI can solve narrow problems and this is called artificial narrow intelligence. For example, a smartphone can use facial recognition to identify photos of an individual in the Photos app, but that same system cannot identify sounds.

B

Black Boxes: We call things we don't understand, "black boxes" because what happens inside the box cannot be seen. Many machine learning algorithms are "black boxes" meaning that we don't have an understanding of how a system is using features of the data when making their decisions (generally, we do know what features are used but not how they are used). There are currently two primary ways to pull back the curtain on the black boxes of AI algorithms: interpretat machine learning (see definition above) and explainable machine learning (see definition below).

C

Chat-based generative pre-trained transformer (ChatGPT): A system built with a neural network transformer type of AI model that works well in natural language processing tasks (see definitions for neural networks and Natural Language Processing below). In this case, the model: (1) can generate responses to questions (Generative); (2) was trained in advance on a large amount of the written material available on the web (Pre-trained); (3) and can process sentences differently than other types of models (Transformer).

Computer Vision: Computer Vision is a set of computational challenges concerned with teaching computers how to understand visual information, including objects, pictures, scenes, and movement (including video). Computer Vision (often thought of as an AI problem) uses techniques like machine learning to achieve this goal.

Critical AI: Critical AI is an approach to examining AI from a perspective that focuses on reflective assessment and critique as a way of understanding and challenging existing and historical structures within AI. Read more about critical A

D

Data: Data are units of information about people or objects that can be used by AI technologies.

Deep Learning: Deep learning models are a subset of neural networks. With multiple hidden layers, deep learning algorithms are potentially able to recognize more subtle and complex patterns. Like neural networks, deep learning algorithms involve interconnected nodes where weights are adjusted, but as mentioned earlier there are more layers and more calculations that can make adjustments to the output to determine each decision. The decisions by deep learning models are often very difficult to interpret as there are so many hidden layers doing different calculations that are not easily translatable into English rules (or another human-readable language).

E

Explainable Machine Learning (XML) or Explainable AI (XAI): Researchers have developed a set of processes and methods that allow humans to better understand the results and outputs of machine learning algorithms. This helps developers of AI-mediated tools understand how the systems they design work and can help them ensure that they work correctly and are meeting requirements and regulatory standards.

> It is important to note that the term "explainable" in the context of explainable machine learning or explainable AI, refers to an understanding of how a model works and not to an explanation of how the model works. In theory, explainable ML/AI means that an ML/AI model will be "explained" after the algorithm makes its decision so that we ca understand how the model works. This often entails using another algorithm to help explain what is happening as the "black box." One issue with XML and XAI is that we cannot know for certain whether the explanation we are getting is correct, therefore we cannot technically trust either the explanation or the original model. Instead, researchers recommend the use of interpretable models.

F

Foundation Models: Foundation Models represent a large amount of data that can be used as a foundation for developing other models. For example, generative AI systems use large language foundation models. They can be a way to speed up the development of new systems, but there is controversy about using foundation models since depending o where their data comes from, there are different issues of trustworthiness and bias. Jitendra Malik, Professor of Compute Science at UC Berkeley once said the following about foundation models: "*These models are really castles in the air, they have no foundation whatsoever.*"

H

Human-centered Perspective: A human-centered perspective sees AI systems working with humans and helping to augment human skills. People should always play a leading role in education, and AI systems should not replace teacher

I

Intelligence Augmentation (IA): Augmenting means making something greater; in some cases, perhaps it means making it possible to do the same task with less effort. Maybe it means letting a human (perhaps teacher) choose to not do all the redundant tasks in a classroom but automate some of them so they can do more things that only a human can do. It may mean other things. There's a fine line between augmenting and replacing and technologies should be designe so that humans can choose what a system does and when it does it.

Intelligent Tutoring Systems (ITS): A computer system or digital learning environment that gives instant and custom feedback to students. An Intelligent Tutoring System may use rule-based AI (rules provided by a human) or use machine learning under the hood. By under the hood we mean the underlying algorithms and code that an ITS is built with. ITSs can support adaptive learning.

Interpretable Machine Learning (IML): Interpretable machine learning, sometimes also called interpretable AI, describes the creation of models that are inherently interpretable in that they provide their own explanations for their decisions. This approach is preferable to that of explainable machine learning for many reasons including the fact that we should understand what is happening from the beginning in our systems, rather than try to "explain" black boxes after the fact.

M

Machine Learning (ML): Machine learning is a field of study with a range of approaches to developing algorithms that can be used in AI systems. AI is a more general term. In ML, an algorithm will identify rules and patterns in the data without a human specifying those rules and patterns. These algorithms build a model for decision making as they go through data. (You will sometimes hear the term machine learning model.) Because they discover their own rules in the data they are given, ML systems can perpetuate biases. Algorithms used in machine learning require massive amounts of data to be trained to make decisions.

> It's important to note that in machine learning, the algorithm is doing the work to improve and does not have the help of a human programmer. It is also important to note three more things. One, in most cases the algorithm is learning an association (when X occurs, it usually means Y) from training data that is from the past. Two, since the data is historical, it may contain biases and assumptions that we do not want to perpetuate. Three, there are many questions about involving humans in the loop with AI systems; when using ML to solve AI problems, a human may not be able to understand the rules the algorithm is creating and using to make decisions. This could be especially problematic if a human learner was harmed by a decision a machine made and there was no way to appeal the decision.

N

Natural Language Processing (NLP): Natural Language Processing is a field of Linguistics and Computer Science that also overlaps with AI. NLP uses an understanding of the structure, grammar, and meaning in words to help computers "understand and comprehend" language. NLP requires a large corpus of text (usually half a million words).

> NLP technologies help in many situations that include: scanning texts to turn them into editable text (optical character recognition), speech to text, voice-based computer help systems, grammatical correction (like auto-corre or grammarly), summarizing texts, and others.

Neural Networks (NN): Neural networks also called artificial neural networks (ANN) and are a subset of ML algorithms. They were inspired by the interconnections of neurons and synapses in a human brain. In a neural network, after data enter in the first layer, the data go through a hidden layer of nodes where calculations that adjust the strength of connections in the nodes are performed, and then go to an output layer.

R

Robots: Robots are embodied mechanical machines that are capable of doing a physical task for humans. "Bots" are typically software agents that perform tasks in a software application (e.g., in an intelligent tutoring system they may of help). Bots are sometimes called conversational agents. Both robots and bots can contain AI, including machine learning, but do not have to have it. AI can help robots and bots perform tasks in more adaptive and complex ways.

S

Self-attention mechanism: These mechanisms, also referred to as attention help systems determine the important aspects of input in different ways. There are several types and were inspired by how humans can direct their attention to important features in the world, understand ambiguity, and encode information.

T

Transformer models: Used in ChatGPT (the T stands for Transformer), transformer models are a type of language model. They are neural networks and also classified as deep learning models. They give AI systems the ability to determine and focus on important parts of the input and output using something called a self-attention mechanism to help.

Training Data: This is the data used to train the algorithm or machine learning model. It has been generated by human in their work or other contexts in their past. While it sounds simple, training data is so important because the wrong data can perpetuate systemic biases. If you are training a system to help with hiring people, and you use data from existing companies, you will be training that system to hire the kind of people who are already there. Algorithms take on the biases that are already inside the data. People often think that machines are "fair and unbiased" but this can be a dangerous perspective. Machines are only as unbiased as the human who creates them and the data that trains them. (Note: we all have biases! Also, our data reflect the biases in the world.)

U

User Experience Design/User Interface Design (UX/UI): User-experience/user-interface design refers to the overa experience users have with a product. These approaches are not limited to AI work. Product designers implement UX/UI approaches to design and understand the experiences their users have with their technologies.

Pati Ruiz and Judi Fusco, "Glossary of Artificial Intelligence Terms for Educators," from the Center for Integrative Research in Computing and Learning Sciences, 2023.