Brief Report

Ethical Decision-Making Guidelines for Mental Health Clinicians in the Artificial Intelligence (AI) Era

Yegan Pillay *

Posted Date: 11 September 2025

doi: 10.20944/preprints202509.1023.v1

Keywords: artificial intelligence; ethics; mental health

 doi:10.20944/preprints202509.1023.v1

Article

Ethical Decision-Making Guidelines for Mental Health Clinicians in the Artificial Intelligence (AI) Era

Yegan Pillay

Ohio University 1; pillay@ohio.edu; Tel.: +1-(740)-593-1000

Abstract

The meteoric rise in generative AI has created both opportunities and ethical challenges in the mental health disciplines namely in clinical mental health counseling, psychology, psychiatry, and social work. While these disciplines have been grounded in well-established ethical principles such as autonomy, beneficence, justice, fidelity, and confidentiality, the exponential ubiquity of AI in society in the past three years has rendered mental health professionals unsure as to how to navigate ethical decision making in the AI era. The author proposes a preliminary ethical framework which synthesizes the code of ethics of the American Counseling Association, the American Psychological Association, the American Medical Association and the National Association of Social Workers which is then organized around five pillars: autonomy and informed consent; beneficence and non-malfeasance; confidentiality, privacy, and transparency; justice, fairness and inclusiveness; and fidelity, professional integrity, and accountability. These pillars are juxtaposed with AI ethical guidelines developed by international organizations, governments, and technology corporations. The resulting integrated ethical framework provides a practical cogent structure that mental health professionals can use when navigating this uncharted terrain. Limitations of the framework and implications for future research are addressed.

Keywords: artificial intelligence; ethics; mental health

1. Introduction

Generative artificial intelligence (AI) as it is known in contemporary society can be credited to a series of steps beginning in 1822 with invention of the computer by Charles Babbage and in the twentieth century by the work of Alan Turing, the British mathematician, computer scientist, and military officer who is considered to be the father of computer science and AI (Grzybowski, et al, 2024) [1]. Although AI has been around for decades, the 21st century has seen an exponential growth in the attention and utilization of generative AI. According to Blinko (2025) [2]ChatGPT which was developed by Open AI, is reported to have gained one million users in the first five days of its launch on September 30, 2022, and has increased from 100 million weekly users in 2023 to approximately 400 million weekly users as of February 2025.

To gauge the attitudes, perceptions, and misperceptions related to the exponential growth in generative AI, the Oliver Wyman Forum [3] conducted two generative AI-specific surveys in June and November 2023 respectively with a sample of roughly 25,000 respondents across 16 countries namely Australia , Brazil, Canada, China, France, Germany, Hong Kong, India, Indonesia, Italy, Mexico, Singapore, Spain, the United Arab Emirates, the United Kingdom, and the United States. One of the survey items asked participants (N=16,033) to respond to the question "Which of the following areas do you think AI will help improve most in the next 30 years?". The findings were as follows: healthcare (41%), transportation (35%), quantum computing (32%), education (32%), media and entertainment (29%), environmental conservation (28%) and energy (25%).

 doi:10.20944/preprints202509.1023.v1

While general health care ranked the highest among the various categories, a notable finding was that 77% (one in three respondents) who have never sought mental health services previously were willing to try generative AI therapy. Moreover, statistical modelling projections by the authors of the Oliver Wyman Forum [3] suggest that generative AI will increase access to 400 million mental health patients worldwide by the year 2030. While these forecasts for the next five years appear promising for the provision of mental health services globally, there remains uncertainty whether AI will transform society positively or lead to negative ethical outcomes. Historically, mental health disciplines in the United States such as clinical mental health counseling, psychiatry, psychology and social work have established ethical parameters that govern the professional practice of their members to ensure the safety of the consumers of mental health services. The rapid pace at which generative AI has impacted society since ChatGPT was developed has left many in society including the mental health gatekeepers without a cogent ethical structure to address the incursion of AI into the mental health spaces.

To address this gap the author examines the various ethical codes of the American Counseling Association (ACA, 2014) [4], the American Psychological Association (APA, 2017) [5], the American Medical Association (AMA, n.d.) [6] and the National Association of Social Workers (NASW, 2021) [7]. In addition, ethics governing AI as advocated by international organizations such as the Organization for Economic Co-operation and Development. (2019) [8], UNESCO. (2021) [9]; by governments such as the US Department of Defense (2020) [10], the UK Government (2020) [11]; by technology corporations such as Google (2018) [12], Microsoft (n.d) [13] , IBM (2018) [14], and academic and non-academic entities such as The Future of Life Institute (2017) [15], The Montreal Declaration (2018) [16], IEEE (2019) [17] and the European Commission High Level Expert Group on Artificial Intelligence (2019) [18] were reviewed with the purpose of proposing a preliminary ethical framework to guide mental health professional practice in this rapidly evolving and uncertain AI landscape.

2. Ethical Principles of Mental Health and AI

The ethical guidelines for the preeminent mental health disciplines are well established with the first code of ethics published for psychologists in 1953 with revisions in 1977, 1992, 2002, 2010 and 2016. This was followed by the adoption of the first ethical guidelines by American Personnel and Guidance Association, the predecessor to ACA, for the for mental health counselors with revisions in 1981,1995, 2004 and 2014. The current version addressed ethical guidelines related to social media, multiculturalism, and diagnosis. The ethical guidelines for social workers were developed in 1979 by the National Association of Social Workers (NASW) with revisions in 1996, 2008, 2013, and 2021. The current version updated the standards on technology, client privacy, and documentation. In 1973 the American Psychiatric Association published 'The Principles of Medical Ethics' which was based on the American Medical Association code of medical ethics that focused on the psychiatric context was updated in 2013 to include confidentiality, boundary violations, and involuntary treatment.

It is evident that the various relevant mental health disciplines codes of practice have a long history with periodic revisions. Although there is no single ethical code for Generative AI, ethical guidelines have recently been formulated by international organizations, governments, technology corporations, and academic and non-academic entities. The principles of the core ethical guidelines of the ACA, APA, AMA, and NASW **[4–7]** namely autonomy and informed consent; beneficence and non-malfeasance; confidentiality, privacy and transparency; justice, fairness and inclusiveness; and fidelity, professional integrity and accountability and the ethical guidelines for AI will be examined for convergent principles with the purpose of informing recommendations for a comprehensive set of integrated ethical guidelines for mental health professionals in the face of the rapid expansion of generative AI in healthcare and specifically for mental health professionals.

 doi:10.20944/preprints202509.1023.v1

2.1. Autonomy and Informed Consent

The mental health disciplines emphasize the client's right to self-determination, respect for their dignity and their right to make choices (ACA, 2014, Section A.1.a; APA,2017, Principle E; AMA, n.d., Opinion 1.1.3 and NASW,2021 Section 1.02**) [4–7**]. Clinicians are required to use culturally and developmental appropriate language to obtain informed consent when conducting assessment, research, teaching, and treatment which includes the goals and techniques, the risks and benefits of therapy, an exploration alternatives and the recognition that informed consent is ongoing and context specific (ACA, 2014 A.2.a.; APA, 2017,Standard 3.10; AMA, 2.1.1; NASW, 2021 Section 1.03**) [4–7**]. According to the IEEE (2019**) [**17] and the OECD (2019**) [**18] AI ethics guidelines indicate that human decision making and human oversight must be preserved and supported. Moreover, AI systems ought to be designed in such a way that they use language that is understood and explainable so that consumers are aware when they are interacting with AI systems so that they can make an informed decision about how their data will be used, stored and transmitted. An important caveat is that the consenting process is done without coercion, and the client can revoke consent at any time.

2.2. Beneficence and Non-Malfeasance

Beneficence and non-malfeasance are ethical principles that guide mental health professionals to promote the well-being of clients while simultaneously abiding by the duty to cause no harm. The ethical codes of the major mental health disciplines address promoting individual well-being, growth, and advocacy for vulnerable populations while simultaneously minimizing treatment interventions or the abuse of power that may cause harm (ACA, 2014 A.1.a, A.4.a; APA,2017 Principle A; AMA, n.d. Principles, I & VIII; NASW, 2017 1.01). Similarly, the AI entities such as IEEE (2019), OECD (2019), EC (2019), IBM (2017), Google(2018), the Future of Life Institute (2017**) [8–17**], advocate to enhance the wellbeing of individuals and society and prevent any unintended harm that may occur. The recommendation by UNESCO (2021) [9] "AI systems must not harm human beings, either through their design or their implementation. The 'do no harm' principle must be upheld as a core value" (https://www.unesco.org/en/artificial-intelligence/recommendation-ethics) encapsulates the sentiment of AI entities.

2.3. Confidentiality, Privacy & Transparency

Ensuring confidentiality, privacy and transparency by mental and AI professionals are prerequisites to informed consent and the development of trust. Mental health professionals are required to protect all client information except in the case of ethical or legal justification (ACA,2014, B.1.b.-B.4.b.; APA, 2017, Standards 4.01–4.07; AMA,n.d. Opinions 3.1.1–3.1.2; NASW,2021, 1.07**) [4–7**] . Mental health professionals ought to allow their clients to control what information they share and with whom the information is shared and clarify confidentiality and its limits. Moreover, the purpose of data collection, especially when AI or digital tools are used in the diagnosis and treatment ought to be addressed as part of the consent process (ACA, 2014,Section B; Section H.; APA,2017, Principle E; AMA,n.d., Opinion 3.3.2; NASW,2021, 1.07 a-n**) [4–7**]. The ethical codes by IEEE (2019), OECD (2019**) [**8,17], requires that AI systems for mental health, health care and social services must ensure that sensitive information e.g. HIPAA is inaccessible to unauthorized parties and is processed with stringent security measures, respects the user's autonomy by obtaining informed consent regarding what data can be collected and that AI systems must be understandable and explainable and users need to be apprised when data is being collected using AI and how decisions are being made for health related issues.

2.4. Justice, Fairness and Inclusiveness

Mental health counselors, psychologists, psychiatrists and social workers by virtue of the ethical codes that guide their practice are expected to avoid discrimination and treat their clients in a manner that that promotes equitable access, social justice and advocacy for marginalized groups in clinical

 doi:10.20944/preprints202509.1023.v1

practice, research and training in the context of social determinants of health (ACA,2014, Preamble, Section A.2.c, C.5, F. 11.c; APA, 2017, Principle D, Ethical Standards 3.01, 9.06; AMA,n.d., Opinion 1.1.2; NASW,2021 **[4–7]**, Preamble, Section 1.05, 6.01, 6.04). The ethical guidelines that govern international, government entities, corporate, academic and non-profit organizations such as OECD (2019), UNESCO (2021), IEEE (2019) **[8,9,17]** , etc. address bias mitigation that promote inclusivity and access to all and non-discrimination policies that promote user engagement in decision making when using AI.

2.5. Fidelity, Professional Integrity and Accountability

The principles of commitment to consumers of mental health service and other professionals, the taking responsibility for one's decisions and actions and the adherence to ethical and moral standard are the building blocks for the establishment of trust in the counseling, psychology, psychiatry and social work disciplines. With the ambiguity and the growth around generative AI, it is imperative that trust is firmly established. Mental health professionals are required to demonstrate the qualities of trustworthiness, avoid deception, report unethical behavior, and seek consultation and supervision, when necessary, in teaching, research and practice (ACA,2014, Preamble, C.3.a.;APA,2017, Principle C; AMA,n.d., Principle II, Opinion 11.2.7; NASW,2021, Preamble, 1.06.4.01**) [4–7]**. Mental health professionals are required to represent their qualifications accurately and maintain competence (ACA,2014, C.4.a-f; APA,2017, 2.03; AMA,n.d., Principle V; NASW,2021, 4.01) [4,7]. The ethical codes of AI as outlined by OECD (2019) and IEEE (2019**) [9,17]** directs AI developers to take into consideration human values and ensure trustworthiness and reliability in combination with creating systems that are transparent, understandable and are fair in both the design and how it is deployed. Moreover, the AI ethical guidelines highlight the imperative that AI developers ensure accountability through a design that comprises of audit, human oversight and provides recourse when harm or an error occurs. The ethical principles guiding mental health professions—counseling, psychology, psychiatry, and social work—have been a dynamic set of parameters that have been developed over an extended period to protect the client in the light of emerging societal and technological changes. The ethical principles emphasize autonomy and informed consent; beneficence and non-malfeasance, confidentiality and transparency; justice, fairness and inclusiveness; fidelity, professional integrity and accountability. Similarly, with the meteoric rise of generative AI, new ethical challenges have emerged, prompting global organizations, government entities, and corporations to develop complementary AI ethics guidelines that stress human-centric values, transparency and explainability, fairness and non-discrimination, privacy and data protection, accountability, safety and robustness, sustainability and social good, and human oversight. The convergence of these frameworks underscores the need for comprehensive ethical guidance to help mental health professionals navigate the integration of AI in clinical, research, and educational settings while safeguarding client rights and well-being.

3. Ethical Framework for Mental Health Professionals

The proposed framework for mental health professionals when navigating the complexities of AI draws from the intersection of the established mental health ethical codes and the recently developed generative AI guidelines. The common ethical principles of the mental health disciplines form the five pillars- namely autonomy and informed consent; beneficence and non-malfeasance, confidentiality and transparency; justice, fairness and inclusiveness; fidelity, professional integrity and accountability-upon which the proposed ethical guidelines have been premised.

3.1. Autonomy and Informed Consent

i. Clinicians must disclose to the client whenever AI is used in their treatment and this disclosure must include AI's capabilities, limitations, potential impacts on their diagnosis, access to treatment, and cost implications.

 doi:10.20944/preprints202509.1023.v1

ii. Clinicians must provide information regarding the type of AI tools that will be used, their impact on the client's treatment, how the data are collected, stored and analyzed and the role and involvement of third parties in the process if relevant.

iii. Clinicians must be willing to allow the client to exercise their right to opt out of AI assisted treatments or decision-making processes and where feasible offer a human-based alternative.

iv. Clinicians must ensure that language used provides clear and understandable details about the use of AI to empower clients to give consent that is fully informed.

3.2. Beneficence and Non-Malfeasance

i. The therapeutic relationship remains central to ethical clinical care and therefore the use of AI must not be viewed as a substitute for human connection but rather as a complementary modality that enhances the therapeutic alliance in assessment, diagnosis and treatment planning.

ii. AI must only be explored as a complementary modality when the clinician determines that they are competent to understand, interpret and explain its results to the client and relevant stakeholders.

iii. Clinicians must select AI tools that are culturally appropriate to minimize the perpetuation of inequities and are reliable, valid, and have evidence-based research support.

iv. Clinicians must use AI tools to promote client well-being and minimize risk rather than justifying or contributing to discriminatory practices which should include algorithms to identify and address ethical concerns or unintended risk.

v. AI tools must adhere to professional and ethical standards and must be evaluated for accuracy and appropriateness on a regularly scheduled basis.

3.3. Confidentiality and Transparency

i. The same confidentiality standards e.g. compliance with HIPAA, federal and state laws, and relevant professional codes including secure data practices, ethical recordkeeping, encryption, and storage protection must also apply to AI tools.

ii. Clinicians must ensure compliance with confidentiality, privacy, and ethical standards, including HIPAA and relevant professional and legal regulations when third-party vendors or AI platforms are used.

iii. Clinicians must be ethically accountable for the use of AI tools and must provide accurate information about AI capabilities and limitations, avoiding misleading claims.

iv. Clinicians should advocate transparency by disclosing how AI models are developed and how algorithms are applied.

3.4. Justice, Fairness, and Inclusiveness

i. Clinicians have a responsibility to ensure that AI systems do not disadvantage individuals or justify discrimination based on marginalized identities.

ii. Clinicians must examine the AI system's function, design, and output to ensure equitable and ethical application and evaluate assessments for cultural bias and fairness.

iii. Clinicians must promote AI systems that enhance justice and safety for all clients and must advocate for inclusive and transparent design features that support ethical obligations of mental health professionals.

3.5. Fidelity, Professional Integrity and Accountability

i. Clinicians should only use AI tools when they have the training and competence to interpret the results responsibly and accurately and fully understand the implications, limitations and the ethical use of AI.

ii. Clinicians must stay informed regarding the best practices and risks regarding emerging tools by staying current with training in AI and digital ethics and in collaboration with AI technologists.

 doi:10.20944/preprints202509.1023.v1

iii. Clinical supervisors must model ethical AI use and guide supervisees in critically evaluating algorithmic tools in their clinical practice.

4. Limitations

This seminal paper offers a preliminary exploration of a framework that provides an ethical structure for mental health professionals navigating the integration of generative AI into their practice, with the caveat that several limitations must be acknowledged. First, generative AI is a rapidly evolving field as is evident in its meteoric rise. Therefore, the proposed framework may have to go through several iterations such as changes with new tools and systems and the potential for accompanying ethical concerns that may emerge in the upcoming years. Secondly, while efforts in this paper were to consolidate human related ethical guidelines with machine related ethical guidelines, this amalgamation may be fraught with practical real-world implications and applications. Third, mental health professionals may not have the necessary technological knowledge to implement AI in their clinical practice effectively and ethically, which speaks to the need for AI to be included in the mental health training curricula. Finally, the proposed framework is preliminary and conceptual and draws from dated ethical codes and emergent AI guidelines. The proposed framework has not been empirically tested and validated in clinical settings. Future research will be needed to evaluate the efficacy of the proposed framework.

5. Conclusions

It is evident that AI has already begun to reshape the societal landscape. If the future mathematical projections by the Oliver Wyman Forum that 40 million more individuals globally will have access to mental health service comes to fruition, it poses both an exciting opportunity and but also adds ethical complexities. While the established flagships for mental health disciplines such as clinical mental health counseling, psychology, psychiatry, and social work already have established ethical guard rails in place, clinical mental health professionals will be traversing uncharted territory in navigating and integrating these established ethical principles with the AI ethical principles advocated by international organizations, governments, technology corporations, and academic and non-academic entities. I advocate that rather than seeing AI as a threat, mental health disciplines ought to explore the opportunities that AI may present with the caveat that we maintain our long-standing ethical compass and identity as mental health professionals while embracing the strengths of generative AI thorough principled and measured ethical lenses.

To this end this paper accentuated the core ethical principles that form the foundation of our ethical practice as mental health professionals which remains crucially relevant and in guiding and navigating our AI integration namely: autonomy and informed consent; beneficence and non-malfeasance; justice, fairness and inclusiveness; fidelity, professional integrity and accountability; and confidentiality, privacy and transparency. Juxtaposed within each of core ethical principles I have weaved in the relevant ethical AI guidelines that resulted in proposed decision-making framework for adapting to the emerging technological realities but still placing the clinician and the client at the center while being advocates for ethical and responsible integration of AI in our respective disciplines.

References

1. Grzybowski, A., Pawlikowska–Łagód, K., & Lambert, W. C. (2024). A history of artificial intelligence. *Clinics in Dermatology, 42*(3), 221–229. https://doi.org/10.1016/j.clindermatol.2023.12.016
2. Blinko (2025) April 14, 2025 . ChatGPT / OpenAI Statistics: How Many People Use ChatGPT? Retrieved from https://backlinko.com/chatgpt-stats?utm_source=chatgpt.com
3. 3.Oliver Wyman Forum. (n.d.). *Oliver Wyman Forum*. Retrieved fromhttps://www.oliverwymanforum.com/index.html

 doi:10.20944/preprints202509.1023.v1

4. American Counseling Association. (2014). *ACA code of ethics*. Retrieved from https://www.counseling.org/resources/aca-code-of-ethics.pdf
5. American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct* (2002, Amended June 1, 2010, and January 1, 2017). Retrieved from https://www.apa.org/ethics/codes
6. **American Medical Association. (n.d.).** *AMA code of medical ethics*. https://code-medical ethics.ama-assn.org
7. National Association of Social Workers. (2021). *Code of ethics of the National Association of Social Workers*. Retrieved from https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of
8. Organisation for Economic Co-operation and Development. (2019). *OECD principles onartificial intelligence*. Retrieved from https://oecd.ai/en/dashboards/policy-areas/ai-principles
9. UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000381137
10. U.S. Department of Defense. (2020). *DoD adopts ethical principles for artificial intelligence*. Retrieved from https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical principles-for-artificial-intelligence/
11. UK Government. (2020). *Ethics, transparency and accountability framework for automated decision-making*. Retrieved from https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making
12. Google. (2018). *AI at Google: Our principles*. Retrieved from https://ai.google/principles/
13. Microsoft. (n.d.). *Responsible AI principles from Microsoft*. Retrieved from https://www.microsoft.com/enus/ai/responsible-ai
14. IBM. (2018). *Everyday ethics for artificial intelligence*. Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
15. Future of Life Institute. (2017). *Asilomar AI principles*. Retrieved from https://futureoflife.org/ai-principles/
16. The Montreal Declaration of Responsible AI (2025) Retrieved from https://montrealdeclaration-responsibleai.com/
17. 17. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (1st ed.). Retrieved from https://ethicsinaction.ieee.org/
18. European Commission High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/ai-allianceconsultation/guidelines#Top