Proceeding of ICIHCS 2025 Symposium: The Dialogue Between Tradition and Innovation in Language Learning
DOI: 10.54254/2753-7048/2025.HT26261

The Ethical Challenges of AI-based Mental Health Interventions: Toward a Layered Accountability Framework

Xinnuo Lei

School of Humanity and Social Science, The University of Hong Kong, Shenzhen, China
122o3oo26@link.cuhk.edu.cn

Abstract. Artificial intelligence (AI) is increasingly prevalent in mental health services, enhancing accessibility by providing immediate support through chatbots and remote platforms, and improving efficiency through automated diagnostics and personalized treatment recommendations. However, this rapid integration also brings numerous ethical controversies, including concerns over data privacy, algorithmic bias, and the potential erosion of human empathy in therapeutic relationships. This paper focuses on seven core ethical issues in AI mental interventions, including privacy protection, informed transparency, fairness and bias, responsibility attribution, autonomy and agency, emotional dependency, and simulated empathy. Existing studies mostly address single dimensions and fail to respond to the multi-stakeholder collaborative ethical challenges. To address this, the paper proposes a "Layered Responsibility Framework" that systematically analyzes the division of responsibilities and ethical constraints across three levels: developers, platform operators, and users. The study highlights that only by promoting clear accountability, transparent design, and institutional coordination can society ensure the sustainable application of AI technology in mental health and safeguard users' psychological safety.

Keywords: Artificial Intelligence, Mental Health, Ethical Challenge, Responsibility Framework

1. Introduction

The way psychological support is obtained, provided, and experienced has changed as a result of the quick integration of artificial intelligence (AI) into mental health care. AI technologies promise to improve accessibility, lower service costs, and close the gap in the delivery of mental health treatment through chatbot-based companionship, automated diagnostic tools, and tailored intervention recommendations. However, this technical development also brings up difficult moral issues that are not sufficiently resolved. Privacy violations, algorithmic bias, unclear responsibility attribution, diminished human agency, emotional dependence, and the appearance of computer empathy are some of these issues. AI-based interventions, in contrast to traditional therapy, frequently function without professional oversight or defined regulatory restrictions. Users may inadvertently provide private psychological information, mistakenly take AI recommendations as medical advice, or grow emotionally dependent on algorithmic agents—particularly in susceptible groups like teenagers. Furthermore, ethical issues are present throughout the AI ecosystem: platform operators affect data governance and user experience, developers impact the training data and underlying logic, and end users frequently engage with systems without sufficient digital psychological literacy. This study suggests a "Layered Responsibility Framework" that divides ethical responsibilities among users, platform operators, and developers in order to address these complex concerns. In order to create a comprehensive governance model that encourages responsible innovation, protects psychological safety, and builds confidence in AI-driven mental health interventions, this study will analyse seven fundamental ethical domains: privacy, informed consent, fairness, accountability, autonomy, empathy, and emotional dependence.

2. Seven core ethical issues

2.1. Privacy and confidentiality

The application of AI in mental health services can efficiently process user data and provide personalized support, but it also poses threats to user privacy, especially regarding sensitive psychological, physiological, and behavioral data. Balancing technical efficiency with privacy protection has become a central ethical issue in AI mental interventions.

Currently, most AI mental intervention tools rely on users continuously inputting large amounts of textual data for interaction and modeling. These data often include emotional states, self-harm ideation, interpersonal conflicts, and even criminal tendencies. Once such sensitive information is illicitly used by third parties or leaked as a result of insufficient platform security measures, users may suffer long-term psychological and social consequences [1]. As Wüller notes, “AI systems rely on data-driven logic, yet mental health data is highly contextualized and difficult to anonymize; thus, 'data anonymization’ strategies have limited effectiveness in this field.” [2].

Moreover, many AI systems are deployed via mobile apps or social media plugins, where users often remain unaware of how their data is collected, used, stored, or shared. Pozzi and De Proost highlight that due to the lack of a “data governance framework,” users unknowingly surrender deep psychological information, thus making them vulnerable to digital exploitation [3]. This “data-deprivation trust” has drawn ethical criticism against the “default consent” model of data collection.

Privacy protection standards for AI mental interventions vary across countries and platforms. Giorgia Pozzi et al. point out that most systems only comply with general privacy policies, lacking refined safeguards specific to sensitive psychological data [3]. Wüller also stresses that AI’s “deep interactivity” in mental health not only collects data but may influence data generation processes, making traditional “inform-consent-use” models inadequate for interactive AI mental interventions [2]. Scholars propose addressing this from both technical and institutional ends: technically enhancing edge encryption, multi-layer access control, and localized data processing [4]. Institutionally promoting the establishment of specialized “psychological data ethics standards,” mandating requirements on data collection, usage, retention, and deletion rights. Artificial Intelligence Ethics in Psychological Support Services recommends that mental health platforms undergo independent ethical review, especially adopting more conservative and transparent data management strategies for vulnerable groups such as adolescents or patients with mental disorders [5].

2.2. Informed consent and transparency

In traditional mental health services, informed consent is a core mechanism to protect client rights, but AI intervention challenges this ethical foundation. AI tools often appear “technologically neutral" but conceal asymmetries of knowledge and design intentions. When using psychological chatbots such as Replika, Wysa, and Woebot, users usually face simplified user agreements and readily agree to services through simplified "click-to-consent" mechanisms, lacking understanding of key operating mechanisms. Wüller observes that users passively accept services without technical comprehension, while AI's ethical façade obscures the operation of power [2].

In clinical contexts, informed consent includes essential information about the service's purpose, methods, risks, benefits, and alternatives. However, when AI performs emotion regulation, cognitive restructuring, or anxiety screening, users find it difficult to clearly identify its "psychological intervention" nature. On one hand, these AI products are often presented as light applications like "daily companionship" or "mood diaries," leading users to mistake them for mere social or lifestyle apps; on the other hand, platforms may downplay their professional psychological functions to evade ethical regulation for data mining and user retention purposes [1]. Consequently, users might unknowingly participate in substantive psychological interventions.

In practice, informed consent is often "formulated" as users must tick lengthy, jargon-heavy agreements upon first login, which ordinary users struggle to truly comprehend. Zidaru et al. note that such "compliance-type consent" mostly serves as platforms' formal responses to regulatory obligations rather than a bi-directional ethical contract based on respect [4]. Even when users read carefully and click "I agree," it does not guarantee their understanding of service nature and data risks, reducing "transparency" to symbolic governance. Kiuchi et al. emphasize that current AI mental health services lack operational standards for informed consent and remain at the level of abstract ethical principles [5]. Ideally, AI platforms should provide sources of decision logic (e.g., data samples, algorithm types, and confidence explanations) and allow users to "visually track" historical conversations to enhance users' understanding and sense of control over personal data.

2.3. Fairness and bias

AI mental health services are often regarded as neutral tools capable of enhancing efficiency, expanding accessibility, and reducing costs. However, research shows that these applications do not equally benefit all groups and may inadvertently worsen disparities in access to mental health care by amplifying structural exclusion of marginalized populations.

Bias in AI systems originates from unbalanced and non-representative data. Mainstream mental health chatbots typically rely on Western-centric corpora such as Reddit, Twitter, or DSM-5 questionnaires, embedding cultural, gender, class, and cognitive model assumptions [6]. For example, Woebot, based on cognitive-behavioral therapy models, assumes users can accurately express emotions and perform rational restructuring, which may not apply to non-native English speakers, users from different cultural backgrounds, or those with limited expressive abilities, potentially causing misinterpretation or misdiagnosis [7].

Algorithmically, AI also displays "mainstream prioritization" in risk prediction, emotion recognition, and intervention recommendations. Depression risk prediction exhibits significantly higher language recognition accuracy for white males compared to females, non-white, and transgender users [8]. Algorithms inherit mainstream social representation logics and value judgments, excluding marginalized groups from technological boundaries.

Developers' "implicit assumptions" further solidify bias structures. Most products assume users are adults, educated, English-speaking, and verbally competent, neglecting psychological needs of disabled, linguistic minorities, and cognitively impaired individuals [9]. Pozzi and De Proost indicate that "cultural middle-class bias" in technical design leads to neglect of impoverished groups in data samples, feature development, and user feedback, obstructing genuine psychological support [3].

Regarding “accessibility,” most AI mental products adopt subscription or tiered payment models, posing barriers for economically disadvantaged populations. Zidaru et al. point out that without public resource intervention, inclusivity goals are difficult to achieve [4]. The Global South and remote regions face challenges from limited internet access, lack of data privacy, and psychological stigma, rendering AI services unable to ensure equity and potentially fostering new forms of “data colonialism” [10].

2.4. Responsibility and accountability

With the widespread use of AI technology in mental health, responsibility attribution and accountability mechanisms become urgent ethical challenges. In traditional counseling models, therapists bear clear legal and ethical obligations, but AI involvement blurs this chain of responsibility. When AI provides inappropriate advice or fails to intervene effectively, who—developers, platform operators, or regulators—should be held accountable? Existing legal and ethical frameworks provide no clear answers [11].

AI mental systems typically involve multiple stakeholders including model developers, technology platforms, deploying institutions, and users, complicating responsibility diffusion and making accountability difficult. In a typical case, a user’s emotional crisis worsens after interacting with an AI chatbot, which only responds with 'Please take care and breathe deeply,” lacking proper referral mechanisms. The platform claims the product is “for emotional companionship only,” developers assert that “the model generates responses automatically based on training data,” resulting in an “unattributable” ethical vacuum [12]. This exposes two issues: first, regulatory systems have significant regulatory gaps. Most countries have yet to establish specialized regulations for AI mental services, lacking pre-ethical review and post-incident accountability mechanisms [13]. Second, functional roles are ambiguous, with AI positioned between health tools and psychological services—neither fulfilling the duties of a therapist nor offering comprehensive counseling support—misleading users to regard it as trustworthy professional support while overlooking ethical, safety, and professional limitations [14].

More alarmingly, some products obscure professional boundaries with rhetoric like “24/7 availability, unbiased, precise analysis.” When users face serious psychological crises, AI systems neither accurately identify risks nor possess emergency handling capabilities [6]. Scholars propose “Ethics by Design,” embedding moral constraint modules into AI mental system architectures to automatically assess the appropriateness and ethical risks of outputs [15]. Yet, most products still rely on user self-judgment, exacerbating accountability difficulties. Data governance problems also highlight accountability dilemmas. Sensitive user psychological privacy data leakage is often evaded by platforms via claims of “voluntary user input” and “data used to optimize algorithms” [16]. This lack of responsibility systems threatens user rights and hinders sustainable AI development in mental health.

2.5. Autonomy and human agency

The broad application of AI in mental health is reshaping humans’ positions in counseling relationships, particularly challenging individual autonomy and agency. Counseling traditionally emphasizes self-exploration and self-determination, but AI’s role as “advisor,” “evaluator,” or even "decision-maker" begins to replace human dominance and intrinsic growth capacities with technological logic.

On one hand, many AI systems rapidly provide emotional judgments, risk alerts, and "coping suggestions" based on users' language or behavioral inputs. This seemingly efficient feedback appears professionally authoritative to users, especially vulnerable or psychologically inexperienced ones. Such reliance may gradually diminish self-awareness and decision-making capabilities when facing emotions and difficulties [11]. Rahsepar Meadi et al., in a systematic review on conversational AI ethics, point out that users often cannot distinguish AI advice from professional clinical opinions, unconsciously relinquishing their own decision rights. On the other hand, AI programs often preset certain "ideal psychological states" and "health models" at design, subtly inducing users toward specific behavioral patterns such as continuous emotion tracking and behavior goals—constituting a form of "soft regulation" of human behavior [4]. For example, some AI mental health apps continually push suggestions according to algorithmic "optimal coping" paths rather than fostering diverse understandings of problems. Though this structural design seemingly promotes "adaptive change," it fails to account for the complexity of human nature and diversity of values.

Therefore, ethical intervention's key is not to prohibit AI use but to establish clear boundaries, ensuring technology enhances human decision-making rather than undermining it.

2.6. Empathy and loss of humanity

AI's expanding use in mental health, especially in emotional support and counseling, marks "empathy" functionality as a technical milestone. AI "empathy" typically refers to affective computing technologies that utilize multimodal data—voice, text, facial expressions—to identify users' emotional states and provide appropriate feedback, creating an illusion of empathetic interaction [5]. However, ethical and psychological reflection reveals that AI-simulated "empathy" conceals its fundamental non-human nature and may produce a dangerous "ethical illusion." Users might mistakenly treat AI as a trustworthy emotional attachment, causing misguidance and risks [17].

First, AI's empathy is entirely algorithm-driven, mechanically recognizing and responding to emotional signals. Although technology can accurately capture multi-layered emotional features such as semantics, tone, and facial expressions, AI lacks genuine emotional experience and subjective perception. It cannot "feel" pain, anxiety, or joy like humans, nor comprehend the complex socio-cultural contexts and personal life histories behind emotions [18]. In other words, AI's so-called "empathy" is only the result of information processing and pattern matching, devoid of emotional understanding [5]. This fundamental difference between technological essence and human emotion means AI empathy lacks intrinsic motivation and value recognition.

Second, the widespread promotion of affective computing in mental health easily generates "ethical illusions," where users mistakenly believe AI feedback possesses empathy quality and humane care equivalent to human counselors [17].

Psychologically, genuine empathy is a product of complex interpersonal interaction, including cognitive empathy (understanding others' feelings) and affective empathy (sharing feelings) [7]. AI only infers emotions based on text and achieves preliminary cognitive empathy relying on statistical models, lacking humans' intuition and inner experience of empathy [5]. This deficiency may come across as mechanistic, emotionally detached, or ineffectual when dealing with complex psychological distress, cultural differences, and value conflicts, failing to meet users' deep psychological needs.

2.7. Over-reliance and emotional dependency

With the popularization of AI companion services such as Replika, adolescent users increasingly rely on these platforms to fulfill emotional needs. These services use natural language processing technology to simulate human communication, attracting users by offering 24/7 availability and non-judgmental interaction. However, problems of over-reliance and emotional dependency have gradually emerged, raising ethical and mental health concerns [19, 20].

Adolescents are in critical stages of personality formation and psychological development, with immature social and emotional regulation abilities. Although AI companions offer prompt responses and superficially warm interactions, they remain algorithmic programs lacking genuine emotional understanding and empathy [5]. Adolescents may perceive AI companions as unconditional supporters, leading to psychological dependence, social avoidance, and weakening of their abilities to cope with social pressures and emotional difficulties [17]. Excessive reliance may impair adolescents' real-world social skills, resulting in feelings of isolation and entrapment in vicious cycles of "virtual dependency" and "social withdrawal" [19].

Additionally, AI companions' personalization algorithms continuously adjust response strategies by analyzing user interaction data, increasing user stickiness and frequency of user engagement [5]. Commercial incentives may foster "technology addiction," causing adolescent users to spend excessive time interacting with AI companions, leading to attention deficits, academic decline, and sleep disorders [20]. These services, leveraging natural language processing and offering constant, non-judgmental user engagement, pose significant risks to adolescent mental health, including emotional dependency, social skill degradation, and real-life isolation, highlighting current AI ethical framework deficiencies and societal negative impacts [17].

3. Ethical governance approach: constructing a layered responsibility framework

AI technology applications in mental health involve multiple actors: model developers, platform operators, and end users, with ethical risks exhibiting layered and synergistic characteristics. Therefore, systematic governance from a "layered responsibility" perspective is necessary to delineate ethical responsibilities across stakeholder levels and build an ethical foundation for sustainable technology development.

3.1. Developer level: source governance embedding ethical design

Ethical risks in AI systems for mental health often originate in the early model construction stage. These systems must assess user emotions, identify potential crises, and propose interventions; thus, their internal logic settings directly affect interaction effectiveness and define ethical boundaries and social impact. Developers must fulfill moral responsibilities from the source.

Morley et al. indicate that sensitive scenarios like mental health require AI systems with high explainability and transparency, posing challenges not only technically but also for accountability and ethical supervision [21]. For instance, algorithms should clearly indicate training data sources, predictive logic, and uncertainty ranges to avoid misleading users or suppressing human judgment.

Furthermore, Fjeld et al. show that AI ethics principles worldwide recommend embedding "non-maleficence," "fairness," and "data governance" into system design from the outset rather than as post-development patches [22]. This "ethics-by-design" approach prevents risks and reduces biases or harms to users.

Data diversity and representativeness are also key to bias prevention. Luxton et al. warn that many mental health AI models rely on English-language and culturally specific datasets, causing mismatch between training data and user group characteristics where AI performs poorly on marginalized groups [23]. Developers should proactively include multicultural, multilingual, and varying ability-level data during development to improve system inclusivity and cross-cultural robustness.

3.2. Platform operator level: crisis prevention and user empowerment

Platform operators act as intermediaries connecting developers and users, controlling user access and managing service delivery. They bear key ethical responsibilities in information transparency, risk management, and user empowerment. McCradden et al. argue that the unique nature of AI mental services makes traditional “one-time authorization” agreements insufficient [24]. Platforms should explore dynamic informed mechanisms, such as embedded prompts during different service stages, helping users gradually understand data handling, system logic, and sensitive functions like screening or intervention, thus achieving true informed consent.

Platforms must establish effective early-warning and referral mechanisms to address potential user crises. Luxton et al. propose “human-AI hybrid intervention” models, wherein AI detection of high-risk signals such as self-harm or despair triggers human review or professional intervention rather than automated interaction continuation [23]. Platforms’ responsibilities encompass not only service provision but also ensuring psychological safety during use.

Due to the extreme sensitivity of mental health data, platforms should subject such data to special governance. Fjeld et al. and Floridi & Cowls agree that platforms must not treat these data as general user data but establish strict collection, usage, and storage norms with independent third-party oversight, ensuring data governance that extends beyond internal corporate policies [22,25]. Ethical responsibility includes not only compliance but respect for user vulnerability and institutional safeguards.

3.3. User level: digital psychological literacy and usage boundaries

Although users are often seen as the “vulnerable party” in ethical governance, their behavior and risk perception affect AI system operational boundaries. In mental health, especially for adolescents or psychologically vulnerable groups, “digital psychological literacy” is crucial. McCradden et al. suggest platforms bear educational responsibilities by providing basic psychological education modules when users first access AI systems, helping distinguish AI advice from professional opinions and understanding boundaries between virtual companionship and real psychological support, preventing unconscious equivalence of AI with therapists [24].

User feedback mechanisms are also vital for AI ethical optimization. Morley et al. stress platforms should provide clear complaint channels and feedback systems, enabling users to voice concerns about misleading information, indifferent responses [21].

4. Conclusion

This study sets out to examine the core ethical challenges associated with the integration of AI into mental health services. While AI technologies promise enhanced accessibility, efficiency, and scalability, they also introduce profound ethical complexities that remain inadequately addressed in both research and practice. Focusing on seven key ethical domains—privacy protection, informed transparency, fairness and bias, responsibility attribution, autonomy and agency, emotional dependency, and simulated empathy—this paper has shown that the ethical risks of AI mental interventions are multi-dimensional and stem from the interplay of technical design, platform operation, and user interaction.

Through detailed thematic analysis and empirical evidence, the study identifies that current ethical dilemmas are not isolated but interwoven, often magnified by the absence of clear governance and multi-stakeholder accountability. For instance, violations of user autonomy are closely linked with design opacity and data asymmetries, while emotional dependency issues often result from platform-level commercial incentives and the lack of user education. These findings indicate that ethical governance cannot rely solely on user compliance or developer goodwill but requires a coordinated and systemic response.

Therefore, this paper proposes a Layered Responsibility Framework, which delineates responsibilities and ethical obligations across three key actor levels: developers are urged to embed ethical principles from the outset, incorporating explainability, fairness, and cultural inclusivity into algorithmic design. Platform operators are expected to strengthen crisis response systems, ensure genuine informed consent, and implement robust data protection protocols. Meanwhile, end users—particularly vulnerable groups—should be empowered through digital psychological literacy education and a clear understanding of AI's capabilities and limitations.

By situating ethical responsibility within this layered architecture, the framework encourages transparency, accountability, and preventive action, rather than reactive governance after harm has occurred.

Ultimately, this study contributes to the field by offering a holistic ethical governance approach that acknowledges the complexity of AI mental health ecosystems. Its insights underscore the urgency of developing multidisciplinary and cross-sector collaborations to co-create ethical, inclusive, and sustainable AI mental health solutions. As AI continues to permeate the landscape of psychological support, only by aligning technological innovation with ethical foresight can society safeguard human dignity, psychological safety, and justice in the digital age.

References

[1] Face-to-Faceless: Exploring the Benefits, Risks, and Ethical Considerations of Using Artificial Intelligence in Therapeutic Contexts, https: //www.diva-portal.org/smash/record.jsf?pid=diva2: 1881801, last accessed 2005/07/15

[2] Wüller, C. A.: Ethical and Practical Implications of Artificial Emotional Intelligence Approached from a Philosophical-Psychological Perspective: The Use Case of Psychiatry, University of Twente. Enschede (2023)

[3] Pozzi, G., & De Proost, M.: Keeping an AI on the mental health of vulnerable populations: Reflections on the potential for participatory injustice. AI and Ethics, 1–11 (2024)

[4] Zidaru, T., Morrow, E. M., & Stockley, R.: Ensuring patient and public involvement in the transition to AI-assisted mental health care: A systematic scoping review and agenda for design justice. Health Expectations 24(4), 1072–1124 (2021)

[5] Artificial Intelligence Ethics in Psychological Support Services: A Scoping Review with Systematic Literature Search, https: //files.osf.io/v1/resources/sztdn_v1/providers/osfstorage/67a01a56cbb9fe3b57445c7c?action=download& direct& version=1, last accessed 2005/07/15

[6] Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V.: Artificial intelligence for mental health and mental illnesses: An overview. Current Psychiatry Reports 21, 1–18 (2019)

[7] Artificial Intelligence in Modern Psychology: A 2025 Guide to Diagnosis, Therapy, and Ethical Innovation, https: //osf.io/preprints/psyarxiv/724qe_v1, last accessed 2005/07/15

[8] Bandawe, S., Roodt, S., & Ruhwanya, Z.: Ethical Limitations of Using AI to Predict and Diagnose Mental Health Disorders Based on Individuals' Social Media Activity. African Conference on Information Systems and Technology 26, (2024)

[9] Falade, O.: Serving Whom? Ethical and Practical Limits of AI Mental Health Chatbots for Marginalized Communities. GRACE: Global Review of AI Community Ethics 3(1) (2025)

[10] Yu, L., & Zhai, X.: Use of artificial intelligence to address health disparities in low- and middle-income countries: A thematic analysis of ethical issues. Public Health 234, 77–83 (2024)

[11] Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N.: Exploring the ethical challenges of conversational AI in mental health care: Scoping review. JMIR Mental Health 12, e60432 (2025)

[12] Luxton, D. D.: Recommendations for the ethical use and design of artificial intelligent care providers. Artificial Intelligence in Medicine 62(1), 1–10 (2014)

[13] Hoose, S., & Králiková, K.: Artificial Intelligence in Mental Health Care: Management Implications, Ethical Challenges, and Policy Considerations. Administrative Sciences 14(9), 227 (2024)

[14] Vale, M. D.: Moral Entrepreneurship and the Ethics of Artificial Intelligence in Digital Psychiatry. Socius 10, 23780231241259641 (2024)

[15] Carvalho, C. A. S.: Ethical challenges of AI-based psychotherapy: The case of explainability. Scenarios: Biannual Journal of Contemporary Philosophy 17(2), 177–199 (2022)

[16] Pandey, H. M.: Harnessing Large Language Models for Mental Health: Opportunities, Challenges, and Ethical Considerations. arXiv preprint arXiv: 2501 (2024)

[17] Balcombe, L.: AI chatbots in digital mental health. Informatics 10(4), 82 (2023)

[18] Fiske, A., Henningsen, P., & Buyx, A.: Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research 21(5), e13216 (2019)

[19] El-Mashharawi, H. Q., Alshawwa, I. A., Salman, F. M., Al-Qumboz, M. N., Abu-Nasser, B. S., & Abu-Naser, S. S.: AI in Mental Health: Innovations, Applications, and Ethical Considerations. International Journal of Academic Engineering Research 8(10), 53-58 (2024)

[20] Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B.: Chatbots and conversational agents in mental health: A review of the psychiatric landscape. The Canadian Journal of Psychiatry 64(7), 456–464 (2019)

[21] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A.: From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics 26(4), 2141–2168 (2020)

[22] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M.: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, Cambridge (2020)

[23] Luxton, D. D., June, J. D., & Kinn, J. T.: Technology-based suicide prevention: Current applications and future directions. Telemedicine and e-Health 17(1), 50–54 (2011)

[24] McCradden, M. D., Baba, A., Saha, A., Ahmad, S., Boparai, K., Fadaiefard, P., & Cusimano, M. D.: Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: A qualitative study. Canadian Medical Association Open Access Journal 8(1), E90–E95 (2020)

[25] Floridi, L., & Cowls, J.: A unified framework of five principles for AI in society. Machine Learning and the City: Applications in Architecture and Urban Design, 535–545 (2022)