

**Ioanna Malandraki, PhD**

Member of NKUA Applied Philosophy Research Lab  
National and Kapodistrian University of Athens  
Greece

## PROTECTING THE SOUL: CONSIDERATIONS ABOUT THE ETHICAL DESIGN OF ARTIFICIAL INTELLIGENCE SYSTEMS FOR PSYCHIATRIC USE

**ABSTRACT:** Can the combination of Artificial Intelligence (AI) and psychiatry ("AI psychiatry") be effective and at the same time ethically permissible? The rapid development of AI and the increasing incidence of mental disorders in recent years call for investigation of this question. Psychiatry, as an inherently sensitive branch of medicine, requires a differentiated approach to diagnosis and treatment. This process is often complex and requires patients, if they are able, to express their emotional states, i.e. the manifestation of their vulnerable side. AI can provide valuable support to psychiatrists, provided that the systems of AI psychiatry are developed on the basis of an ethical design, as it emerges through its alignment with a clear and universally accepted regulatory framework. Consequently, it can be argued that AI psychiatry is undoubtedly being developed to facilitate the physician's profession with its assistance, but due to the vulnerability of the field, it also leads to ethical concerns that lead to ethical concerns. In particular, questions arise about maintaining confidentiality between patient and psychiatrist, and protecting the privacy and rights of patients.

**KEYWORDS:** ethics, bioethics, artificial intelligence, medicine, psychiatry, mental health.

### I. INTRODUCTION

Several years ago, modern society may have seemed like a figment of imagination. Nowadays, individuals can use their mobile phones to monitor their sleeping habits, receive news tailored to their interests, interact with household appliances, learn a foreign language through specialized applications, receive entertainment recommendations, and even engage in discussions with AI on various topics, from washing delicate garments to making important life decisions such as quitting a job. The worldwide rapid evolution of AI across diverse domains of human life is being observed.

Living in this society, my interest in the phenomenon of AI is growing due to the constant conquest of new levels of technology. AI is affecting more and more aspects of our lives, from social media to law and medicine. This trend is compelling contemporary individuals to seek involvement in all evolving events. AI can have a significant impact on people's lives, leading to various concerns. It is important to consider the potential consequences of AI's influence.

In recent years, the rate of mental disorders has steadily increased globally (GBD 2019 Mental Disorders Collaborators 137). Mental illnesses are currently among the top ten most prevalent diseases that impede the harmonious functioning of individuals. The potential increase in their incidence may be a result of the psychological, social, and economic impact of the COVID-19 pandemic, according to a study in The Lancet Psychiatry (GBD 2019 Mental Disorders Collaborators 137).

The integration of AI in psychiatry is a part of the broader context of AI usage in medicine. In fact, this approach promises to improve patient outcomes: psychiatric AI solutions have been implemented and are functioning under supervision. Specifically, AI is currently being used under supervision for the purpose of directly detecting symptoms of mental disorders and initiating the most effective treatment to counteract them at an early stage. Advanced algorithmic and computational techniques, including automated natural language processing and machine learning algorithms, are used to examine individuals' mental states during diagnosis, monitoring, and treatment in psychiatry (Rocheteau 54-55). In addition, as in any application of medical AI, in the specific field of psychiatry, AI contributes to secretarial tasks such as scheduling and interpersonal communication with patients and their families (Lekadir et al. 13).

The combination of AI and psychiatry raises ethical considerations due to the rapid development of AI and the increase in the rate of mental disorders in recent years. Ethical considerations that need to be explored include the protection of confidentiality between patient and psychiatrist, privacy, individualization, and the rights of mentally ill.

## II. APPLICATIONS

At this moment, AI in psychiatry contributes significantly to overall patient care. AI assists in diagnosing, monitoring, treating, and managing related administrative and nursing issues.

During the diagnostic phase, AI is utilized for faster detection of symptoms at an early stage. To achieve a reliable diagnosis, psychiatrists must conduct an individual examination and avoid relying solely on previous experience of symptoms. Wearable applications have been developed to assist physicians in achieving reliable diagnoses. The development of wearable applications is particularly advantageous for diagnosis because patients can easily record numerical and mood data in their homes (Rocheteau 55).

At the next phase, at the monitoring phase, AI is used to continuously monitor patient health data. Fitness trackers, also known as sports watches for measuring physical activity, can help individuals monitor their health by tracking metrics such as heart rate, anxiety levels, and sleep duration (Rocheteau 55). Research has shown that individuals with severe mental health conditions find these devices helpful for monitoring their health (Rocheteau 55). Smartphones can also provide valuable information about an individual's location and social activities providing information which can be useful, for example, in examining a patient on the spectrum of bipolar disorder who is experiencing a manic phase (Rocheteau 55). Vocal data, such as pitch, rhythm, and volume, are biomarkers of many psychiatric illnesses, including depression and anxiety (Rocheteau 55).

The impact of AI on patient treatment is primarily notable for its predictive capabilities, especially in anticipating the effects of specific psychiatric medications (Rocheteau 55). Deep learning techniques enable the prediction of treatment outcomes based on medication responses, allowing for personalized treatment plans that reflect the patient's needs. Additionally, mobile phone applications for chatbot-psychotherapists have been developed. Examples of such chatbots are Wysa, Woebot, Limbic, RiseUp, and Mindspa. These chatbots are used by individuals seeking immediate solutions to their problems but who do not wish to communicate with a human being in the conventional manner (Rocheteau 55). It is important to note that

these apps explicitly state that they should not be considered authorities or used as an alternative to treatment administered by a practitioner (Rocheteau 55-56).

In medical AI, including psychiatric AI, AI systems perform administrative and nursing roles. Initially, these systems manage patient appointment scheduling and using historical data can suggest an ideal date for the next session (Lekadir et al. 13-14). They can also predict possible patient no-shows based on previous activity, thereby enabling early intervention by psychiatrists (Lekadir et al. 13-14). These systems use extensive databases to estimate waiting times and suggest appropriate times to patients to avoid negative experiences that could affect the overall quality of their healthcare (Lekadir et al. 13-14). Additionally, AI systems can monitor financial transactions to detect and prevent fraudulent activities (Lekadir et al. 14). Moreover, patient's queries can be resolved by providing advice based on their electronic health records (Lekadir et al. 14). Finally, AI systems can be used as consultants to assist physicians in making challenging decisions and navigating ethical dilemmas (Aderson et al., "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson" 188-189) –this is the vision of the Machine Ethics research project, proposed by Philosophy professors Michael Anderson and Susan Leigh Anderson (Anderson & Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent" 15).

### III. BENEFITS, DRAWBACKS, AND REFLECTIONS

From a positive standpoint, some have noted that the absence of certain human elements in psychiatric AI systems can lead to more accurate diagnoses and treatment for patients. Specifically, the potential beneficial impact for patients could address issues such as fatigue, stress, and distraction that may affect psychiatrists during the day (Ray et al. 4). In other words, psychiatric AI systems eliminates those human factors from assessments.

Also, it is important to note that many patients still resist seeing a psychiatrist. This refusal stems from the difficulty individuals experience in sharing their troubles and from the social stigma (Ray et al. 4). Research has investigated the impact of online support groups for patients with depression and has found that those with severe depressive disorders who choose to isolate themselves socially can benefit from the support group due to the anonymity it provides (Houston, Cooper, and Ford 2062-2063 & 2066-2067). Psychiatric AI encourages open expression. Therefore, it can be inferred that disclosing symptoms to a clinician through AI is likely to be more comfortable for patients providing a non-judgmental space for them (Ray et al. 4).

Additionally, psychiatric AI has the potential to identify new biomarkers that indicate the presence of mental disorders. The objective is to identify individuals who are at risk of developing mental disorders and predict their health status after taking certain drugs to ensure appropriate care (Terra et al. 2). This will reduce costs associated with mental health care and will provide access to psychiatric mental health systems for all individuals.

By supporting the ideal healthcare standard – evidence-based prevention, diagnosis, and treatment of diseases, patient-centered personalized, quality care (Lee and Yoon 11) – psychiatric AI systems promote mental healthcare in remote or developing areas, supporting open access. This is particularly important in areas where people have limited healthcare access (Ray et al. 4).

On the other hand, it is important not to overlook the potential drawbacks of psychiatric AI. One of the primary concerns is that the field of psychiatry places greater

emphasis on establishing a therapeutic alliance between the patient and the clinician (Stubbe 402). This connection is typically established by addressing the patients' emotions and behaviors through observation. The current state of psychiatric AI development lacks the necessary human elements, such as empathy and compassion, to effectively cultivate this relationship. Psychiatric AI systems not only eliminate human factors such as fatigue and stress, but also human factors such as empathy and compassion from assessments. The assistance provided to patients with mental illness can be restricted without these elements (Ray et al. 4).

Additionally, the long-term use of psychiatric AI systems may create a dependency relationship; continuous interaction with the system may pose a risk of fostering addictive behaviors in patients. The use of such systems to facilitate the expression of patients' emotions raises concerns about emotional comfort. Patients may develop attachment to these systems and exhibit addictive behaviors, leading them to avoid seeking treatment from a psychiatrist (Terra et al. 7). The lack of protection of medical confidentiality and medical privacy is necessary. It is important to note that patient safety may not always take precedence (Terra et al. 7).

Additionally, dependence on non-deterministic programming can lead to unpredictability in AI decision-making. Psychiatric AI systems may make incorrect diagnoses due to a lack of training in identifying all mental illnesses. AI bias can result from the inability of algorithms to process untrained data (Terra et al. 7) or from misclassification and labeling of data by developers (Blasimme and Vayena 706). Related to this risk limitation, current machine learning is unable to track changes in mental disorders over time (Terra et al. 7).

There are common considerations for all AI applications, particularly in the medical field. The main concern is liability. Using AI to treat patients in medicine presents a potential for errors. In such cases, where does responsibility rest among the system, psychiatrist, and programmers (Rocheteau 56)? Secondly, there is the issue of opacity. The process for extracting decisions from an AI system is unclear. In conflicts between the psychiatrist's decision and the system, which holds greater reliability? In cases where there is a discrepancy between the decision made by the psychiatrist and the decision made by the system, it is unclear whether the verdict of the system should be relied upon due to its non-explicable nature (Rocheteau 56). Additionally, how can patient data protection be ensured amidst the integration of AI into psychiatric practices? Patient confidentiality and privacy protection are important considerations. It is crucial to ensure that AI systems uses patient data in a secure manner.

#### IV. THE REQUEST FOR ETHICAL DESIGN

Codes of ethics are essential for both AI and psychiatry. Firstly, it is crucial to design AI systems in a way that ensures their ethical beliefs. Secondly, psychiatry must consider psychiatric ethics to practice properly, as there are ethical complexities specific to mental health care. In the diagnosis and treatment of a patient's mental illness, psychiatrists must be both scientifically competent and morally responsible for their decisions regarding emerging issues related to the general rights of the mentally ill.

Bioethics is an appropriate field for examining ethical issues that arise in psychiatric AI due to its interdisciplinary nature. The four fundamental principles of biomedical ethics provide an ethical framework for addressing these questions. The principle of autonomy defends an individual's right to freely choose actions and

decisions that do not impinge on the freedom of others (Flynn). The principles of beneficence and non-maleficence aim to increase well-being and avoid harm, respectively (Flynn). The principle of justice upholds the fair distribution of medical good, ensuring that each person receives their just due (Flynn).

The ethical considerations arising from the partnership between AI and psychiatry are diverse. The primary question is how to design psychiatric AI systems to ensure their use aligns with the ethical principles in psychiatry and ultimately with biomedical ethics. The development of such systems must consider moral accountability, the incorporation of values into the system, and democratization.

The initial concern pertains to the difficulty in attributing moral and legal accountability. This challenge arises when an AI system causes harm through wrongful actions that could potentially violate the principles of beneficence and non-maleficence. This issue is primarily rooted in AI bias and the opaque inner decision-making process of AI systems. The intricate structure and operation of these systems give rise to the black box problem. Due to this structure, it is challenging to avoid biased decisions that individuals choose while relying on the trust placed in these systems. Furthermore, the black box problem prevents human-supervisors from correcting incorrect decisions, as the algorithmic reasoning guiding AI systems to their conclusions is beyond of our realm of knowledge, rendering us incapable to rectifying such decisions.

To address the issue and prevent the consequences of the responsibility gap, it is necessary to ensure accountability. Accountability is inextricably linked to responsibility, which involves documenting proposals or actions and ethical causality activating the demand for transparent AI systems (AI transparency) (Diakopoulos 197). Fundamentally, accountability refers to the creation of a relationship between of two entities, where one entity substantiates the information provided and the other one accepts or rejects the substantiation (Kroll 183). The realization of accountability provides a foundation for clarifying the attribution of legal and ethical context of AI entities, because in cases where the system has made an incorrect action "[...] while the objects of recordkeeping are generally machines, software, or algorithms, the entity being held answerable must be a moral worthy of the ascription of responsibility (Kroll 190)." Otherwise, if it is determined that the AI system did not engage in any wrongdoing, it should be investigated through an audit process to attribute accountability to the user(s), developer(s), and/or those who decided and use the system.

Transparency is crucial for ensuring the safety of AI systems. It is crucial because people seek reliability, which can only be achieved through understanding how these systems behave. In order to achieve this understanding, transparency in parallel with explainability are required, videlicet the human-supervisor must have access to a description of the system's behavior (transparency) and a causal explanation of its decision-making process (explainability) (Diakopoulos 204-205).

The second issue relates to the examination of the moral and legal values that are legitimate to be instilled in AI systems to fulfill their objectives set in each application. This issue pertains to upholding the principles of autonomy, beneficence, and non-maleficence. Within this framework, the protection of the patient's personal data and the protection of the patient's cognitive privacy are discussed, as the preservation of confidentiality in psychiatry is a fundamental principle for the sealing of an ethical treatment;

Whatever the psychiatrist has been told by the patient, or has noted during examination or treatment, must be kept confidential unless the patient relieves the

psychiatrist from this obligation, or to prevent serious harm to self or others makes disclosure necessary. In these cases however, the patient should be informed of the breach of confidentiality. (Declaration of Hawaii/II)

The observance of privacy entails the protection and the respect for individuals, as privacy is an indisputable right of every human being. Computer science professor Stuart Russell argues that AI systems can and should incorporate privacy as a fundamental feature (Russell 71). It is important to note that Stuart Russell understands and clarifies that the majority of AI applications require the inclusion of personal data in order to function properly (Russell 71). Nevertheless, the professor believes that this action should not result in retreat and cessation of claim privacy, because "learning algorithms can operate on encrypted data using the techniques of secure multiparty computation, so that users benefit from pooling without compromising privacy in any way (Russell 71)."

In addition, cognitive privacy must be protected. This type of privacy follows the progress of machine learning and Neuroscience; "privacy on the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses (Sadkhan 269)." For instance, in computational psychiatry, which is defined as the study of mental disorders and symptoms through Neurosciences, limiting its conclusions to the biological characteristics of individuals, it is imperative to ensure cognitive privacy due to the access and recording of brain neural activity with the ultimate goal of its analysis.

One mechanism for achieving patient's privacy, particularly in a European context, is to align system objectives with the criteria outlined in the General Data Protection Regulation (GDPR) or the EU AI Act.

The GDPR sets out detailed requirements for companies and organizations on collecting, storing and managing personal data. It applies both to European organizations that process personal data of individuals in the EU, and to organizations outside the EU that target people living in the EU (Data protection under GDPR).

Another mechanism for achieving patient's privacy is cybersecurity. This mechanism is essential for safeguarding AI systems in medicine against cyber-attacks that could potentially lead to patient fatalities (Lekadir et al. 24-25);

On September 9 2020, a woman died during a cyber-attack on a hospital in Düsseldorf, Germany. The woman was in a critical condition and about to be treated when hackers disabled the computer systems of the hospital. Unable to avert the attack, medical staff had to transfer the woman to another hospital, but the help came too late and the woman died (Kiener).

The third issue aims to provide equal access to AI for all individuals to allow for expanded participation. The issue at hand is the legitimacy of granting everyone a voice in the development, organization, and usage of AI systems and whether patients should be involved in these proceedings. The final step towards ethical design of systems, while respecting the principle of justice, brings to the fore that the fair distribution of the benefits of AI in society is of paramount importance, that this step can be achieved by opening AI to more developers and physicians for its design, and

that the demand for transparent AI systems is urgent (Rubeis, Dubbala and Metzler 4).

The democratization of AI highlights the need for transparent AI systems, as it is suspected that the implementation of AI systems without transparency in the context described above could be counterproductive. This raises the question of whether the involvement of patients in the development of AI systems is ultimately in their best interest. Due to the ease of access to medical AI systems, there is a risk of over-diagnosis, wherein a disease may be detected that would not cause harm or death even without treatment (InformedHealth.org). There is also a risk of increased medicalization, which can lead to social control, as certain problems are treated as medical conditions (Conrad and Bergey 105). In fact, the bias and opacity of AI systems, as well as the programming of these systems that fails to derive error probability estimates for recommendations, may increase psychological harm to patients under these circumstances.

## V. CONCLUSION

Mental, as well as physical, health is invaluable. Psychiatric AI can assist in ensuring ethical standards and upholding the principles of biomedical ethics, but only if it is designed to meet these parameters. Otherwise, the use of psychiatric AI can increase the vulnerability of mental health. To address this issue, psychiatric AI systems must incorporate these principles to ensure ethical and effective use. Systems must be designed, implemented, and operated in strict adherence to ethical principles during development under a regulatory framework, or/and psychiatrists must be provided with continuous guidance on ethical conduct aligned with these principles. Ethics can contribute to the advocacy, promotion, and improvement of mental health through the ethical design and ethical use of psychiatric AI. "Whether we have to fear AI technology depends on how it is developed and by whom it is used. In itself it is neutral (Anderson et al., "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson" 193)." Ethical considerations aren't an option but a prerequisite in designing AI for psychiatric purposes.

## REFERENCES

1. Anderson, Michael and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." AI Magazine 28, no. 4 (2007): 15-26. Web. 24 April 2024.
2. Anderson, Michael, Susan Leigh Anderson, Alkis Gounaris and George Kosteletos. "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson." Conatus – Journal of Philosophy 6, no. 1 (2021): 177-202. Web. 24 April 2024.
3. Conrad, Peter and Meredith Bergey. "Medicalization: Sociological and Anthropological Perspectives." International Encyclopedia of the Social & Behavioral Sciences. Ed. J. D. Wright. Amsterdam: Elsevier, 2015. 105-109. Web. 24 April 2024.
4. "Data protection under GDPR." Your Europe. 7 June 2022. Web. 24 April 2024.
5. "Declaration of Hawaii/II." World Psychiatric Association. Web. 24 April 2024.
6. Diakopoulos, Nicholas. "Transparency." The Oxford Handbook of Ethics of AI. Eds. M.D. Dubber, F. Pasquale, S. Das. USA: Oxford University Press, 2020. 197-214. Print.
7. Flynn, Jennifer. "Theory and Bioethics." The Stanford Encyclopedia of Philosophy. 25 November 2020. Web. 24 April 2024.

8. GBD 2019 Mental Disorders Collaborators. "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019." *The Lancet Psychiatry* 9, no. 2 (2022): 137-150. Web. 24 April 2024.
9. Houston, Thomas K., Lisa A. Cooper, and Daniel E. Ford. "Internet Groups for Depression: A 1-Year Prospective Cohort Study." *American Journal of Philosophy* 159, no. 12 (2002): 2062-2068. Web. 24 April 2024.
10. InformedHealth.org [Internet]. "What is overdiagnosis?" Institute for Quality and Efficiency in Health Care. 20 April 2017. Web. 24 April 2024.
11. Kiener, Maximilian. "'You may be hacked' and other things doctors should tell you." *The conversation*. 3 November 2020. Web. 24 April 2024.
12. Kroll, Joshua A. "Accountability in Computer Systems." *The Oxford Handbook of Ethics of AI*. Eds. M. D. Dubber, F. Pasquale, S. Das. USA: Oxford University Press, 2020, 181-196. Print.
13. Lee, DonHee and Seonh NoYoon. "Application of Artificial Intelligence-Based Technologies in the Healthcare Industry: Opportunities and Challenges." *International Journal of Environmental Research and Public Health* 18, no. 1 (2021): n. pag. Web. 24 April 2024.
14. Lekadir, Karim, Gianluca Quaglio, Anna Tselioudis Garmendia and Catherine Gallin. "Artificial Intelligence in healthcare. Applications, risks and ethical and societal impacts." *European Parliamentary Research Service. Scientific Foresight Unit PE 729.512* (2022): n. pag. Web. 24 April 2024.
15. Ray, Adwitiya, Akansha Bhardwaj, Yogender Kumar Malik, Shipra Singh, and Rajiv Gupta. "Artificial Intelligence and Psychiatry: An overview." *Asian Journal of Psychiatry* 70 (2022): n. pag. Web. 24 April 2024.
16. Rocheteau, Emma. "On the role of artificial intelligence in psychiatry." *The British Journal of Psychiatry* 222, no. 2 (2023): 54-57. Web. 24 April 2024.
17. Rubeis, Giovanni, Keerthi Dubbala, and Ingrid Metzler. "'Democratizing' artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term." *Front. Genet* 13 (2022): 1-11. Web. 24 April 2024.
18. Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Penguin Books, 2019. Print.
19. Sadkhan, Sattar B. "Cognitive and the future." 2018 International Conference on Advance of Sustainable Engineering and its Applications (ICASEA). IEEE, 2018. 269-270. Web. 24 April 2024.
20. Stubbe, Dorothy E. "The Therapeutic Alliance: The Fundamental Element of Psychotherapy." *Focus – The Journal of Lifelong Learning in Psychiatry* 16, no. 4 (2018): 402-403. Web. 24 April 2024.
21. Terra, Mohamed, Mohamed Baklola, Shaimaa Ali, and Karim El-Bastawisy. "Opportunities, applications, challenges, and ethical implications of artificial intelligence in psychiatry: a narrative review." *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery* 59 (2023): 59-80. Web. 24 April 2024.