Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

OPEN ACCESS

doi.org/10.51891/rease.v10i8.15298

BEYOND BLACK BOXES: INTERPRETABLE AI FOR ENHANCED RISK ASSESSMENT AND ETHICAL DECISION-MAKING IN FORENSIC PSYCHIATRY

Yasmin Vitória Carvalho de Castro[1]
Kelly Yumi Morii[2]
Igor Araújo Santos[3]
Afrânio Côgo Destefani[4]
Vinícius Côgo Destefani[5]

ABSTRACT: The increasing adoption of artificial intelligence (AI) in forensic psychiatry has sparked discussions about its potential to revolutionize risk assessment, diagnosis, and treatment. However, the use of 'black box' AI models, which lack transparency and interpretability, has raised significant ethical concerns. This narrative review explores the current state of AI in forensic psychiatry, with a focus on developing interpretable AI models for enhanced risk assessment and ethical decision-making. The article underscores the importance of considering social and environmental factors alongside neurobiological data in AI-based predictions and discusses AI's legal and ethical implications in forensic contexts. The review concludes by emphasizing the need for interdisciplinary collaboration and responsible evaluation of AI models before widespread adoption in high-stakes decision-making processes within forensic psychiatry and criminal justice.

Keywords: Forensic Psychiatry. Artificial Intelligence. Risk Assessment. Ethics, Professional. Decision Making.

INTRODUCTION

Forensic psychiatry plays a crucial role in the intersection of mental health and the legal system, with risk assessment being a central component of clinical practice (1). The advent of artificial intelligence (AI) has opened new possibilities for enhancing the accuracy and efficiency of risk assessment, diagnosis, and treatment in forensic psychiatry (2). However, using "black box" AI models that lack transparency and interpretability has raised ethical concerns (3). This narrative review aims to provide an overview of the current state of AI in

[1]Centro Universitário São Lucas (Porto Velho/RO).
[2]Centro Universitário São Camilo.
[3]Centro universitário UniFG. e-mail:
[4]Santa Casa de Misericórdia de Vitoria Higher School of Sciences - EMESCAM. Santa Luíza – Vitória ES Brazil, Molecular Dynamics and Modeling Laboratory (DynMolLab.
[5]Molecular Dynamics and Modeling Laboratory (DynMolLab). Santa Luíza – Vitória – ES – Brazil.

forensic psychiatry, focusing on developing interpretable AI models for enhanced risk assessment and ethical decision-making. The article also discusses AI's legal and ethical implications in forensic contexts and emphasizes the need for responsible evaluation of AI models before their widespread adoption.

METHODOLOGY

A comprehensive literature search was conducted using Scopus, Web of Science, PubMed, and ScienceDirect databases. The search terms included combinations of "artificial intelligence," "machine learning," "forensic psychiatry," "risk assessment," "ethics," and "interpretability." Relevant articles published in English between 2020 and 2024 were selected for inclusion in the narrative review. The articles were analyzed for their content, and the findings were synthesized to provide an overview of the current state of AI in forensic psychiatry, focusing on interpretable AI models and ethical considerations.

RESULTS

The current state of AI in forensic psychiatry

1.1. Risk assessment and prediction of violent behavior

AI techniques, particularly machine learning algorithms, have shown promising results in predicting violent behavior and recidivism in forensic populations (1). Studies have demonstrated that AI models can outperform traditional methods in terms of accuracy and efficiency (2). However, many of these models rely on "black box" approaches that lack transparency and interpretability (3).

1.2. Diagnosis and treatment

AI-based diagnostic tools, especially those utilizing neuroimaging data, have been developed with high accuracy and efficiency. AI interventions, such as chatbot-based therapy and virtual reality exposure therapy, have shown early signs of effectiveness in forensic mental health treatment (2). However, the integration of these tools into clinical practice remains limited due to ethical and legal considerations.

Interpretable AI models for enhanced risk assessment

2.1. The importance of interpretability

Interpretable AI models, also known as "glass box" models, provide transparency in their decision-making processes, allowing clinicians and legal professionals to understand the factors contributing to risk predictions (3). This transparency is crucial for ensuring fairness, accountability, and trust in AI-based risk assessments (1).

2.2. Incorporating social and environmental factors

Externalist accounts of psychiatric disorders emphasize the importance of considering social and environmental factors alongside neurobiological data in AI-based predictions (4). Interpretable AI models that incorporate these factors can provide a more comprehensive and contextualized understanding of an individual's risk profile (4).

2.3. Practical implications for AI model design

The development of interpretable AI models for risk assessment in forensic psychiatry has practical implications for data collection, processing, and the selection of machine learning methods (4). Researchers and developers must ensure that training data is representative, unbiased, and includes relevant social and environmental variables. The choice of machine learning algorithms should prioritize interpretability without compromising predictive performance (3).

Legal and ethical implications of AI in forensic psychiatry

3.1. Informed consent and privacy

The use of AI in forensic psychiatry raises concerns about informed consent and privacy, mainly when dealing with sensitive mental health and criminal history data (1). Clinicians and researchers must ensure that individuals are fully informed about the nature and purpose of AI-based assessments and that their data is securely stored and protected (5).

3.2. Fairness and bias

AI models trained on biased or unrepresentative data can perpetuate or amplify existing disparities in the criminal justice system, particularly along racial lines (6). Ensuring fairness and mitigating bias in AI-based risk assessments is a critical ethical obligation for researchers and practitioners in forensic psychiatry (1).

3.3. Accountability and legal admissibility

The admissibility of AI-based risk assessments in legal proceedings is an ongoing debate, with concerns about the transparency, reliability, and validity of these tools (2). Establishing clear standards for the development, validation, and use of AI in forensic psychiatry is essential for ensuring accountability and legal admissibility (3).

DISCUSSION

The findings of this narrative review highlight the potential of AI to revolutionize risk assessment and decision-making in forensic psychiatry, while also underscoring the ethical and legal challenges associated with its implementation. The development of interpretable AI models that incorporate social and environmental factors alongside neurobiological data represents a promising avenue for enhancing the transparency, fairness, and contextualized understanding of risk profiles (4). However, the responsible development and deployment of these models require careful consideration of ethical principles, such as informed consent, privacy, fairness, and accountability (1,6).

The legal admissibility of AI-based risk assessments in forensic contexts remains a contentious issue, with ongoing debates about the transparency, reliability, and validity of these tools (2). Establishing clear standards and guidelines for the development, validation, and use of AI in forensic psychiatry is crucial for ensuring its legal admissibility and promoting trust among clinicians, legal professionals, and the public (3).

CONCLUSION

AI has the potential to transform risk assessment and decision-making in forensic psychiatry, offering more accurate, efficient, and contextualized approaches to predicting

violent behavior and recidivism. Developing interpretable AI models incorporating social and environmental factors is a promising direction for enhancing transparency, fairness, and ethical decision-making in forensic contexts. However, the responsible development and deployment of these models require careful consideration of ethical principles and legal implications. Interdisciplinary collaboration among clinicians, researchers, ethicists, and legal professionals is essential for navigating the complex landscape of AI in forensic psychiatry and ensuring its beneficial impact on individual lives and society.

REFERENCES

1. TORTORA l, meynen g, bijlsma j, tronci e, ferracuti s. Neuroprediction and a.i. in forensic psychiatry and criminal justice: a neurolaw perspective. Front psychol [internet]. 2020 mar 17;11. Available from: https://www.frontiersin.org/article/10.3389/fpsyg.2020.00220/full

2. AGGARWAL nk, jain a. Neuroethics and neurolaw in forensic neuropsychiatry: a guide for clinicians. Behavioral sciences and the law. 2024;42(1).

3. GARRETT bl, rudin c. Interpretable algorithmic forensics. Proceedings of the national academy of sciences [internet]. 2023 oct 10;120(41). Available from: https://pnas.org/doi/10.1073/pnas.2301842120

4. STARKE g, d'imperio a, ienca m. Out of their minds? Externalist challenges for using ai in forensic psychiatry. Front psychiatry [internet]. 2023 aug 24;14. Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1209862/full

5. TORTORA l. Beyond discrimination: generative ai applications and ethical challenges in forensic psychiatry. Front psychiatry [internet]. 2024 mar 8;15. Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2024.1346059/full

6. HOGAN nr, davidge eq, corabian g. On the ethics and practicalities of artificial intelligence, risk assessment, and race. J am acad psychiatry law [internet]. 2021 sep;49(3):326–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/34083423