Journal of Psychiatry and Cognitive Behaviour
Reich J. J Psychiatry Cogn Behav 9: 207.
www.doi.org/10.29011/2574-7762.000207
www.gavinpublishers.com
Review Article

Can Artificial Intelligence Aid in Understanding Forensic Psychiatry Case Law? An Exploratory Study

James Reich MD

Volunteer Professor of Psychiatry, Department of Psychiatry Stanford Medical School, USA.

*Corresponding author: James Reich, Volunteer Professor of Psychiatry, Department of Psychiatry Stanford Medical School, USA.

Citation: Reich J (2025) Can Artificial Intelligence Aid in Understanding Forensic Psychiatry Case Law? An Exploratory Study. J Psychiatry Cogn Behav 9: 207. DOI: https://doi.org/10.29011/2574-7762.0000207

Received Date: 06 September, 2025; **Accepted Date:** 18 September, 2025; **Published Date:** 22 September, 2025

Abstract

Background: Large language models of artificial intelligence (AI) are used in many areas of medicine. This study investigated whether AI would be helpful to the forensic psychiatrist practitioner in understanding case law. **Method:** The American Academy of Psychiatry and the Law (AAPL) runs a review course for Forensic Psychiatry Boards. AAPL develops multiple choice practice questions for participants. Six of these questions were presented to two commonly used AI programs: ChatGPT and Perplexity. In addition the quality of AI references was compared against a similar search using PubMed. **Results:** ChatGPT correctly answered two of the six questions while Perplexity correctly answered four of the six questions. ChatGPT and Perplexity did not give the same answer on four of the six questions. In the two cases where both agreed the answer was correct. Both AI systems explained their reasoning. PubMed supplied four times more peer reviewed and up to date references than either AI system. **Conclusions:** Although both AI programs answered the correct questions at above chance level neither was at a level that could be relied upon without further investigation. Agreement between the two systems improved their accuracy. Regardless of agreement AI results need to be checked against other sources.

Keywords: Artificial intelligence; ChatGPT; Forensic science; law; Psychiatry; Perplexity.

Introduction

The goal of this article is to try to understand the usefulness of AI for the forensic psychiatrist in the area of case law. This will be done first by examining accuracy of two AI systems on multiple choice forensic case law questions and then by assessing the value of AI cited peer review articles against a PubMed search. The hypothesis was that AI search engines would be a useful tool for forensic psychiatrists and provide correct answers and useful references. AI refers to computer systems capable of performing tasks that typically require human intelligence, such as learning from experience, problem-solving, and decision-making. Large language models (LLMs), like OpenAI's Generative Pre-Trained Transformer (GPT) series, are a subset of AI designed to understand and generate natural language. These models are pre-trained on vast amounts of text data, enabling them to perform tasks such as text completion, summarization, and question-answering with high fluency and coherence.

AI has entered the medical field. Specialties such as radiology, pathology, dermatology, and cardiology are already using AI for image analysis [1-3] and studies have been done to determine whether it could improve diagnostic reasoning in internal medicine, emergency room and family practice physicians [4].

Large data set analysis has been used in forensic psychiatry. There are models to estimate suicide risk [5], prediction of sexual offenders reoffending [6], prediction of violent reoffending after release from prison [7] and prediction of repeat domestic violence [8]. However, there are significant potential limitations of AI. In a review of AI to generate clinical summaries many errors were noted. These errors included variability where different answers were given to the same question; "sycophancy" where AI tailors

answers to the perceived user expectation in the prompt; and "complete the narrative" errors where additional information was added [9]. Given the increasing use of AI with potential relevance forensic psychiatry the current study was designed to give information about its use in the area of forensic case analysis.

Methods

Test questions given to AI consisted of the first six questions board practice questions from the 2016 American Academy of Psychiatry and the Law (AAPL) forensic psychiatry board review course. AAPL was contacted for permission to use of these questions in this study. Board questions are designed to test a board applicant's knowledge of forensic psychiatry. These questions were designed by a committee of senior forensic psychiatrists to mimic board review questions. Although not all board review practice questions are on forensic case law, a majority are. These six questions were all about case law. The questions were asked two highly used AI programs, Chat GPT and Perplexity. Their answers were compared to the board review correct answers on the answer sheet which in this case was considered the correct answer. The overall accuracy of the AI models was compared to chance answering.

As a second test of the usefulness of AI I checked the references cited in the AI programs against references found by a similar query in PubMed.

Results

For the board review questions Perplexity gave the first answer correctly. For that first question it indicated that it lacked information on several of the cases in the multiple-choice set, although it gave a brief description of each case. This phenomena of the program noting it had incomplete data occurred a few times in these tests. It could mean that the case was mentioned somewhere in its data set but the AI program did not have the full case in its data set or that the program did not have access to the information at all. The second question was missed because the relevant case was not in its data set. It did report that the data was missing from its data set. Perplexity answered the next three questions correctly. On the final question it indicated that none of the answers was correct, an incorrect answer. Perplexity gave an explanation of its reasoning for all its answers in what seemed a logical way. Perplexity's correct answer rate of 66% was higher than the 25% rate by chance.

The same questions were posed to ChatGPT. The program gave correct answers to two of the six questions. It gave an explanation of the reasoning in all of its answers to all six questions. ChatGBT explained its reasoning in what seemed a logical way, however, four of the answers were wrong. Overall ChatGPT's correct score of 33% of the answers correct was above the chance rate of 25%. The program indicated on some, but not all of its wrong answers that it lacked complete data on the question (Table 1).

Question asked	Possible answers	Correct answer	Perplexity	Chat GBT
1. The following cases all involve a not identifiable victim EXCEPT:	A. Petersen v. State of Washington B. Naidu v. Laird C. Lipari v. Sears D. Tarasoff v. Regents	D	D, but indicated a lack of information on several cases.	C
2. The following cases all involve confidentiality issues EXCEPT:	A. Doe v. Roe B. Jaffee v. Redmond C. Tarasoff v. Regents D Naidu v. Laird E. Petersen v. State of Washington	D	Cannot answer without further information.	E
3. The following cases all involve duty to protect EXCEPT:	A. Lipari v. Sears B. Rock v. Arkansas C. Jablonski v. U.S. D. Naidu v. Laird E. Tarasoff v. Regents	B	B	B
4. The following cases all involve informed consent EXCEPT:	A. Clites v. Iowa B. Whalen v. Roe C. Canterbury v. Spence D. Kaimowitz v. Michigan DMH	B	B	B

5. The following cases all involve malpractice EXCEPT:	A. Clites v. Iowa B. Canterbury v. Spence C. In re Lifschutz D. Naidu v. Laird E. Roy v. Hartogs	C	C	E
6. The following cases all involve convicted prisoners' rights EXCEPT:	A. Barefoot v. Estelle B. Estelle v. Gamble C. Vitek v. Jones D. Baxstrom v. Herold E. Farmer v. Brennan	A	Indicated that none of the answers were correct.	D

Table 1: Forensic Board Review Questions from an American Academy of Psychiatry and the Law 2016 board review course asked ChatGPT and Perplexity compared to correct answers.

Looking at both programs together they agreed correctly on two of the six questions and disagreed on four of the six questions. The two questions that the AI programs agreed were both indicated to be correct answers by the correct answer.

To test another aspect of AI, the references provided by Perplexity and CHATGPT were compared to those found in PubMed. PubMed provided more scientifically sound and up to date references. Both AI programs produced mostly references for the general population such as newsletters. PubMed cited refereed scientific journals. Overall PubMed Cited four times as many refereed published articles as did the AI programs.

Discussion

Looking at our results for correctness of the answers to forensic case questions each AI model had significant limitations. The answers for Chat GPT were only slightly above the chance of 25% at 33% correct. The Perplexity answers were a bit higher at 66% which is better but not high enough to be relied upon.

When we examine the answers both AI programs got correct it becomes more interesting. The AI programs are do not have identical software, have not been programed with the same data set, and therefore do not respond to questions in an identical fashion. If they are responding above chance level, although not with complete accuracy, two programs coming to the same answer would indicate a higher probability of correctness than just one as the probability of error would be lower. This would indicate the possibility that when two or more AI programs agreed the chances of a correct answer would be greatly improved. An agreement by three AI programs would indicate a higher level of probability above chance or about 2% in the case of a four choice question (.25 x.25 x.25).

Important to assessing the value of AI answers is the question of whether the AI systems had a sufficient data base. If the data fed into the program is incomplete or without context, the output will be as well [10]. Other researchers have speculated that with incomplete date the results could be biased [11, 12]. AI programs can only cite information in their data base and expanding that base is expensive. Scientific findings are published after the AI cutoff date of data collection and are often highly relevant. In addition, if articles cited by our AI searches are any indication, scientific articles were less frequent in the AI training data. In addition, Generative AI models can amplify existing biases and create synthetic data that may not accurately reflect real-world conditions. AI capabilities must be thoroughly evaluated before widespread adoption in any given area [13].

In our findings comparing the references provided in AI searches to a similar search in PubMed the references produced by AI were largely newsletters or non-peer reviewed publications. They also were less recent than PubMed. This is likely due to the cutoff date of their most recent data entry being less recent than articles available on PubMed. AI is a powerful tool but can only work with the data entered into its system. The massive data entry to create or update an AI data set is expensive. Therefor not all relevant data for a given specialty such as forensic psychiatry may be present in a given data set. Experts should recognize this issue and use AI whose data set is more in line with their professional needs and need to be aware of the limitations of the programs data sets that they do use. The forensic psychiatrist using the system will need to address the question of whether the program itself had a sufficient data base for the questions being asked.

In addition to the data base issue is the problem that, at times, AI just makes up a reference or data, commonly referred to as an "hallucination" [14]. The hallucinations that are sometimes produced by AI searches represent a significant hurdle. It is unlikely that the user will easily recognize a hallucination. For example, one study indicated that the general public can only identify an AI generated fake picture 61% of the time [15]. However, there are some appropriate precautions that can be taken. The user should ask is whether the program was designed for the specific question

in mind. For example, a general AI program may not have specific data related to medical or forensic questions. A second issue is the care with which the question was designed. The better designed and thought through the question, the better the answers AI will give. A possible safeguard might to be to use two or more different AI programs for the same question as AI programs are not identically programed and do not hallucinate in synchrony with each other.

Related to the above issue is that when the AI program had insufficient access to the full data but some access to information it could give an incorrect answer with a reasonable rationale. This could mislead an AI user that the incorrect answer was actually correct. There is no a way to determine whether an answer given by AI is correct just by reading or looking at it. Other references need to be consulted. Other important issues related to AI are that if precautions are not taken, sensitive patient information can wind up on the web [16] and uncritical use of AI can increase malpractice risk [17].

This begins to give us an overall idea of AI's usefulness in forensic psychiatry. When the AI program had full information in its data set it could give the correct answer. AI programs also give their reasoning. There are significant limitations. One was failure to give the correct answer. Another is that AI gave some answers that looked correct, including the reasoning, but were not correct. A third is that different AIs may not answer the same question the same way. Fourth, the AI data base may not be up to date or contain key relevant data. There are limitations to this analysis. As AI programs are constantly updated their results may change over time. This represents a challenge to replication. The analysis was valid at the time it was done.

Conclusions

The integration of AI into forensic psychiatry offers potential benefits but currently has significant limitations. These include: wrong answers; reasoning that appears correct but is not; lack of an adequate data base for forensic psychiatry and/or a data base that is not up to date; and confidentiality issues. Although solutions to some of these issues may be developed in the future AI is not currently a standalone technique for working with forensic case law.

Financial Support

There was no financial support for this project.

Conflicts of interest

There are no conflicts of interest.

I consulted the Committee on Publication Ethics (COPE) website and did not find any deviation from their guidelines. As there were no human subjects involved in the manuscript there were no ethical issues involving human subjects.

I would like to acknowledge the helpful comments of Steven Hyler, MD.

References

1. Bates DW, Auerbach A, Schulam P, Wright A, Saria S (2020) Reporting and implementing interventions involving machine learning and artificial intelligence. Ann Intern Med 172: 137-144.
2. Bates DW, Levine D, Syrowatka A, Kuznetsova, M, Craig. KJT, et al. (2021) The potential of artificial intelligence to improve patient safety: a scoping re- view. NPJ Digit Med 4: 54.
3. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25: 44-56.
4. Goh E, Gallo R, Hom J, Strong E, Weng Y, et al. (2024) Large Language Model Influence on Diagnostic Reasoning A Randomized Clinical Trial. JAMA Netw Open 7: e2440969.
5. Haroz EE, Rebman P, Goklish N, Garcia M, Suttle R, et al. (2024) Performance of Machine Learning Suicide Risk Models in an American Indian Population. JAMA Netw Open 7: e2439269.
6. Yu R, Molero Y, Långstrom¨ N, Fanshawe T, Yukhnenko D, et al. (2022) Prediction of reoffending risk in men convicted of sexual offences: development and validation of novel and scalable risk assessment tools (OxRIS). J Crim Justice 82: 101935.
7. Fazel S, Chang Z, Fanshawe T, Långström N, Lichtenstein P, et al. (2016) Prediction of violent reoffending on release from prison: derivation and external validation of a scalable tool. Lancet Psychiatry 3: 535-543.
8. Yu R, Molero Y, Lichtenstein P, Larsson H, Prescott-Mayling L, et al. (2023) Development and Validation of a Prediction Tool for Reoffending Risk in Domestic Violence. JAMA Netw Open 6: e2325494.
9. Goodman KE, Yi PH, Morgan DJ (2024) AI-Generated Clinical Summaries Require More Than Accuracy. JAMA 331: 637-638.
10. Tortora L (2024) Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry. Front Psychiatry 15: 1346059.
11. Hogan NR, Davidge EQ, Corabian G (2021) On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race. J Am Acad Psychiatry Law 49: 326-334.
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science. Science 366: 447-453.
13. Starke G, D'Imperio A, Ienca M (2023) Out of their minds? Externalist challenges for using AI in forensic psychiatry. Front Psychiatry 14: 1209862.
14. Dinis-Oliveira RJ, Azevedo RMS (2023) ChatGPT in forensic sciences: a new Pandora's box with advantages and challenges to pay attention. Forensic Sci Res 8: 275-279.
15. Lu Z, Huang D, Bai L, Qu J, Wu C, et al. (2023) Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

16. Oliva A, Grassi S, Vetrugno G, Rossi R, Della Morte G, et al. (2021) Management of Medico-Legal Risks in Digital Health Era: A Scoping Review. Front Med 8: 821756.

17. Mello MM, Guha N (2023) ChatGPT and Physicians' Malpractice Risk. JAMA Health Forum 4: e231938.