AI-Assisted Deception and the Emerging Challenge of LLMs in Forensic Psychiatry

Jason G. Roof, MD

Generative artificial intelligence (AI), including the large language model ChatGPT, has introduced potential new opportunities and challenges to the practice of forensic psychiatry. These powerful AI-based tools may offer substantial benefits in administrative tasks, report generation, and record summarization yet simultaneously present areas for further consideration, such as aiding evaluees in feigning psychiatric symptoms. Additional ethics and legal considerations exist regarding privacy, bias within AI models, and the introduction of fabricated or misleading AI-generated content into forensic assessments. Legislative efforts, privacy safeguards, and professional guidelines essential for responsible AI use are being developed. Forensic psychiatrists are uniquely positioned to influence responsible AI integration through education, advocacy, and development of best practices within psychiatry.

J Am Acad Psychiatry Law 53(2) online, 2025. DOI:10.29158/JAAPL.250022-25

Key words: artificial intelligence; large language models; malingering; forensic psychiatry

In this month's issue of The Journal, we are invited by Dr. Gershan and colleagues[1] to consider the future of forensic psychiatry as we face the present reality that generative artificial intelligence may help forensic evaluees malinger psychiatric symptoms. We find ourselves in the era of "AI-assisted deception."

LLMs in Forensic Psychiatry

Large language models (LLM), such as Open AI's Generative Pre-trained Transformer (GPT) Series, Google's Bard, Meta's LlaMA, and Anthropic's Claude are freely available to anyone with a connection to the Internet. LLMs may be endlessly useful in the practice of academic forensic psychiatry. One may research, summarize published articles for a high-yield review, brainstorm topics for academic writings or presentations, develop curriculum outlines, or handle repetitive administrative tasks with only a few keystrokes.

LLMs are but one example of what may be referred to as a foundation model. By using large amounts of data, including text, images, audio, and video, and a process of self-supervised learning, an application may be molded for use in a variety of purposes. ChatGPT, one of the most popular LLMs and the technological focus of the Gershan *et al.*[1] article, is trained on a vast collection of texts to understand and generate human-like communication. A foundational model used to assist a pulmonologist looking for lung cancer could instead be trained on many millions of chest images to provide a helpful alert when an abnormality is detected.[2] LLMs are already highly integrated into our medical practices and medical education.[3] Functions already being assisted by LLMs include administrative tasks, knowledge augmentation, medical education, and medical research.[4]

In their pilot analysis, Gershan and colleagues[1] explore whether ChatGPT can facilitate malingering in a forensic evaluation setting, particularly in the area of feigning psychotic symptoms to diminish or evade legal responsibility. Their preliminary findings suggest an emerging threat in forensic evaluation, that of increasingly more sophisticated AI-based tools that may be utilized to malinger mental health disorders. We may find some solace that our skills and experience in detecting incongruity in presentation, corroborating collateral data, or interpreting contradictory historical

Published online May 27, 2025.

Dr. Roof is a clinical professor of Psychiatry, Associate Training Director, Division of Psychiatry and the Law, Department of Psychiatry & Behavioral Sciences, University of California, Davis, Davis, CA. Address correspondence to: Jason G. Roof, MD; E-mail: jgroof@ucdavis.edu.

Disclosures of financial or other potential conflicts of interest: None.

records will still serve us well in our evaluations. We cannot, however, remain complacent and ignore the ever-increasing impact such technology will have on our profession and our society.

Readers may believe they have successfully avoided AI-assisted software. Note that Microsoft 365's Copilot acts as an AI assistant within its Office apps and can generate smart summaries of lengthy emails and propose a better way to write an email to a coworker. Zoom and Teams offer AI-assisted summaries of meetings. Gmail offers "smart replies" to others emails, providing a quick suggested response. Google searches offer an "AI Overview," which summarizes relevant searched sources for your convenience. Grammarly, a commonly used software plugin for word processors, offers "tone checks" that can rephrase entire written paragraphs for any number of goals. AI-driven chatbots (utilizing conversational AI) are frequently first-line customer support options for your pharmacy, bank, or insurance providers. Social media AI algorithms consider billions of posts to attempt to screen out harmful content and to better target you for increased engagement on their platform.

Power Tools Without Instruction Manuals

As forensic psychiatrists, we are trained to assess malingered psychiatric symptomatology on its own or as it applies to various medicolegal questions involving capacity and competency. We are now confronted with evaluees and attorneys who are potentially equipped to fabricate or embellish mental health symptoms in ways that may be more difficult to detect. We are also confronted with our own access to an increasingly available array of powerful tools that are being rapidly developed and deployed without sufficient regulation or input by clinical medical providers who may be asked to utilize these tools in their practices.

For a moment, consider the lure of conducting a forensic assessment utilizing a "clinical AI system for social behavior verification" (Ref. 5, p 1), which dissects, digests, and interprets an evaluee's medical and mental health records as well as the content of your forensic interview. Once the data are internalized and integrated into the model, you have the opportunity to directly interact with the data, asking it any questions you like. You might utilize such a tool as a predictive model for violence toward self or others if the preexisting dataset were reported to be sufficient. You could ask the model to generate a partial or even full forensic report. It would be possible to ask this tool for constructive feedback on your own forensic opinion. You could feed an opposing forensic expert's opinion into the dataset to look for potential weaknesses. If you did so, you would need to consider how you would explain this process to an attorney or to finders of fact on the stand.

The opportunity for forensic psychiatrists to interact with and automatically summarize vast amounts of collateral records and synthesize meaningful responses, including forensic reports, is alluring. Such decisions come with clear benefits as well as unknown and unforeseeable consequences. We neglect opportunities to consolidate our knowledge, develop critical thinking skills, and integrate knowledge to be used for future cases or testimony when such critical thought is offloaded. Early career forensic psychiatrists and trainees may inadvertently sidestep supervisory opportunities where deep nuanced discussions may occur. Explaining reasoning behind decisions for finders of fact must go further than reciting data output from an LLM. Additionally, integrations of such tools will lead to an indelible integration of this tool in one's practice and a risk for increased future dependence on its availability.

While being mindful of our own implicit and explicit biases, we are also asked to consider how we might address bias inherently embedded in the LLM itself. LLMs are trained on vast bodies of data, which undeniably will contain the longstanding biases, prejudices, and systemic inequalities within our larger society.[6] An LLM's output can reflect or even amplify such biases, particularly in the absence of safeguards or oversight. LLMs cannot transcend the limitations of their dataset and embody the best and worst of us all.

Legal Considerations

Great effort is occurring in implementing safeguards and further defining the limits of AI in health care settings.[7] California AB 3030 seeks to better legislate how care involving AI must be disclosed to the patient.[8] California Senate Bill 1120, also known as the Physicians Make Decisions Act, strives to avoid AI decision-making without human oversight as it relates to utilization management.[9] Concerningly, at this time, there are no comprehensive, AI-specific laws that directly govern the use of AI in clinical or forensic settings. A preexisting network of institutional guidelines and local, state, and federal laws provide some early rudimentary guideposts, such as the protection of sensitive health information in the

privacy and security standards of the Health Insurance Portability and Accountability Act (HIPAA) standards.[10,11]

Consideration has been given to the troubling implications of submitting sensitive medical information into a public-facing LLM. Increasingly, institutions are obtaining and deploying institution-specific LLMs and providing warnings against use of public-facing LLMs. Institution-specific LLMs are utilized only by a specific site or held within a HIPAA-compliant, encrypted cloud environment to prevent sensitive data from leaving its secure infrastructure. Physicians violating these privacy laws place themselves at increased risk for civil and criminal exposure as well as professional disciplinary action.

Guidelines and Safeguards

Professional organizations and individual authors have suggested guidelines related to a variety of evaluation-related technological processes, including forensic evaluations conducted with videoconferencing software.[12,13] Few have offered guidelines or best practices related to the evaluator's use of AI or how to prepare for potential LLM use by a forensic evaluee or their attorney.

Collaboration and coordination with one another within our field of forensic psychiatry will be key in educating ourselves and helping others to learn about what effect AI may have on medicine and forensic psychiatry. Guidelines and best practices related to the mental health and forensic use of AI-enhanced resources should be heavily influenced by forensic psychiatrists and by the American Academy of Psychiatry and the Law. We are uniquely positioned on a national stage to provide meaningful and practical advice in this complex and rapidly evolving use of technology.

Responsible use of any new powerful tool originates from a place of familiarity with its capabilities and limitations, and we must do the work to engage and influence use of LLMs. This will need to begin at the level of medical student education and continue throughout our careers.[14] Our efforts may involve advocacy for responsible usage within clinical and forensic settings as well as providing clearheaded, practical education at state and legislative levels. AI is here and deeply integrated into our personal and professional lives. Its influence will only continue to grow.

In closing, it is time to ask ourselves what can be gained by freeing up our minds and time to safely complete tasks we may find less fulfilling or too complex. We need to consider what would be lost in moving toward a professional model where critical thinking is more likely to be automated and outsourced. We must develop comprehensive guidelines to include responsible use, limitations, and safeguards as they relate to training, clinical practice, and forensic evaluation. Without addressing these opportunities within our profession, we may find others are more than happy to write the rules. In such a case, an AI-assisted malingerer may be the least of our concerns.

References

1. Gershan SA, Schoenfeld E, Grabb DJ. A pilot analysis investigating the use of AI in malingering. J Am Acad Psychiatry Law. 2025 Jun; 53(2):000–000
2. Lenharo M. An AI revolution is brewing in medicine. What will it look like? Nature. 2023; 622(7984):686–8
3. Gordon M, Daniel M, Ajiboye A, *et al*. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024; 46(4):446–70
4. Omiye JA, Gui H, Rezaei SJ, *et al*. Large language models in medicine: The potentials and pitfalls. Ann Intern Med. 2024; 177 (2):210–20
5. Anibal J, Gunkel J, Awan S, *et al*. The doctor will polygraph you now. Npj Health Syst. 2024; 1(1):1
6. Fisher CE. The real ethical issues with AI for clinical psychiatry. Int Rev Psychiatry. 2025; 37(1):14–20
7. Anderson AJM, Paulette C, Sarata AK, Wells N. Artificial intelligence (AI) in health care [Internet]; 2024 Dec 30. Available from: https://www.congress.gov/crs-product/R48319. Accessed March 18, 2024
8. California Legislature. California Assembly Bill 3030 [Internet]; 2024 Mar 21. Available from: https://legiscan.com/CA/text/AB3030/id/2965727/California-2023-AB3030-Amended.html. Accessed March 25, 2025
9. California Legislature. California Senate Bill 1120 [Internet]; 2024 Sep 9. Available from: https://legiscan.com/CA/text/SB1120/id/2927303. Accessed March 25, 2025
10. Rezaeikhonakdar D. AI chatbots and challenges of HIPAA compliance for AI developers and vendors. J L Med & Ethics. 2023; 51(4):988–95
11. Li J. Security implications of AI chatbots in health care. J Med Internet Res. 2023; 25:e47551
12. Miller TW, Clark J, Veltkamp LJ, *et al*. Teleconferencing model for forensic consultation, court testimony, and continuing education. Behav Sci & L. 2008; 26(3):301–13
13. Shore JH, Yellowlees P, Caudill R, *et al*. Best practices in videoconferencing-based telemental health April 2018. Telemed J E Health. 2018; 24(11):827–32
14. Ötleş E, James CA, Lomis KD, Woolliscroft JO. Teaching artificial intelligence as a fundamental toolset of medicine. Cell Rep Med. 2022; 3(12):100824