Journal of MultiDisciplinary Evaluation
Volume 19, Issue 45, 2023

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

AI for Evaluators: Opportunities and Risks

Aaron W. Kates
Effect X

Kurt Wilson
Effect X

Background: The emergence of widely available and applicable artificial intelligence (AI) raises ethical, practical, and professional concerns for professional evaluators. The authors explore potential answers to emerging questions as to how evaluators can engage AI in an effective and responsible way.

Purpose: Advance the conversation around AI technology and its integration into professional evaluation practice.

Setting: Not applicable.

Intervention: Not applicable.

Research Design: Not applicable.

Data Collection and Analysis: Not applicable.

Findings: Authors explore two main use cases for AI: namely, proposal writing and evaluation design drafting. We also discuss four challenges for evaluators engaging with AI: The proliferation of the digital environment with excess output, market disruption and the emergence of new roles, the so-called "alignment problem", and the challenge to evaluators' to use their agency. Moving forward, the authors recommend evaluators familiarize themselves with AI technology, use it transparently, think critically about the effects of AI on their work, and use perspective when considering the potential ramifications of this new tool.

Keywords: *artificial intelligence; evaluation; technology; evaluation practice*

Background

In the late 1990s when my (author Aaron Kates's) father joined my grandfather's law firm, he brought a Windows 95 computer with him. My grandfather regarded the machine as a "fad"—nothing to be taken seriously. My father, on the other hand, used it to become more efficient—allowing him to work with fewer admin staff.

We are all familiar with how computers and smartphones transformed work and productivity. We now stand at a new threshold with another technology that is predicted to further transform the way we work: artificial intelligence (AI).

In this article, we hope to take a sober and practical look at AI from the perspective of professional evaluators. We will examine how we might use it, how it might transform the nature of our tasks, what threats it might pose to our field, and what evaluators might do to protect themselves and their society from potential adverse effects of this emerging technology.

What is AI?

This section provides an overview of AI as context to understand the implications for evaluation: what it is, where it has come from, and how it works.

What is AI? The European Commission's high-level expert group on artificial intelligence offers the following definition:

> Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions. (2019, p. 6)

To put it in terms that are less nuanced (and therefore less accurate) would be to describe AI as a system that appears to think at a level at least approaching, and in some cases surpassing, that of human intelligence.

While AI products have been in use for a number of years in narrow applications (think text suggestions in gmail), they have begun to play an increasingly important role in the public imagination. The first iteration of AI products to hit the mainstream, gaining mass usership, was developed by a company called OpenAI. The first product, DALL-E, was released to the general public in September of 2022 (OpenAI, 2022a). This product creates images based on a text prompt entered by the user. Images produced can either be photorealistic or mimic an artistic style. The second major advancement, ChatGPT, also by OpenAI, was released to the general public in November of 2022 (OpenAI, 2022b). This system operates as a chatbot in a web browser. It can answer questions, write lengthy pieces of text, and have lifelike conversations with the user.

GPT stands for "generative pre-trained transformer," which describes the program's underlying structure. Rather than being a top-down programmed AI (the traditional way of coding), this system was created using machine learning. This particular form of machine learning is called a large language model. The model assigns a probability to a string of words, allowing the system to predict what should come next in a sentence (Brants et al., 2007). In these systems the AI is given a massive amount of text (essentially the entire internet) and trained to predict the next word of a new sentence based on the statistical relationship observed between words in the massive training corpus. The system is then fed queries and rewarded by programmers to train the AI toward the correct type of responses. Therefore, ChatGPT and similar AI tools are effective prediction models for language. The ability to answer a range of queries in a convincing and factual manner flows from the accuracy of their predictions.

While this method of creating the chat model has produced stunning results, it has also led to some difficulties, chief among them the veracity problem. The web page of OpenAI is clear about this—emphasizing that answers are not always factually correct, so one must use caution (OpenAI, 2022b). To explain why, Dr. Gary Marcus (2022), neuroscientist and AI theorist, explains that ChatGPT is at its most fundamental level a text-prediction engine, and the system is therefore an expert mimic. However, this does not grant the system awareness of the exterior world. This results in the production of language that looks plausible without necessarily being true.

Additionally, since this is an emergent system, rather than one that has been created in a traditional engineering fashion, it is difficult to look under the hood and understand how it is operating (Castelvecchi, 2016; Durán & Jongsma, 2021). This leads many in the tech safety and ethics sphere to raise concerns that need to be taken seriously, because real-world risks are already coming to light. For example, the first suicide involving

interaction with an AI chatbot was reported just two weeks before this writing (Walker, 2023).

Potential Uses for AI in Evaluation

How should professional evaluators engage these developments? Because it has emerged so recently, there is little writing about the application of AI in evaluation. Currently there are no scholarly articles in the evaluation literature (e.g., *American Journal of Evaluation*, *New Directions for Evaluation*, *Journal of MultiDisciplinary Evaluation*) that directly address the topic. The only piece we could find is a blog post by Bruce (2023) that highlights some takeaways for evaluators, including the ability to code and, in the future, the ability to analyze data.

We will try here to give an overview of the opportunities AI presents for evaluation at the moment. The central concept is that the large language models (i.e., ChatGPT) are a useful set of tools for anyone—including evaluators—whose job involves generating text. Here are a few use cases to highlight the point:

Use Case 1: Proposal Writing

Imagine having an assistant who works almost for free—with good style, grammar and punctuation. At worst, this assistant sometimes makes overstatements and makes you sound a little bit too good. You can provide this assistant a stack of ten evaluation proposals you have written and ask them to answer a new RFP in your style.

Now imagine that this assistant can complete the task in under a minute—creating a proposal perfectly tailored to the RFP and mimicking your style. The extensive time you once needed to develop a proposal is reduced to just conveying the task and reviewing the output. This is the promise of AI. While the current iteration of ChatGPT doesn't allow upload of personal reference files, existing iterations are useful for speeding the process of developing a first draft—and movement toward this more refined scenario is progressing quickly as third-party vendors develop such capabilities.

This scenario obviously introduces new problems. Dramatically decreasing the time required to submit a proposal will likely exacerbate longstanding problems with evaluation contracting. Evaluators may find themselves responding to far more RFPs, only to have their hit rate decrease dramatically. Conversely, funders may find themselves flooded with responses to even the smallest RFPs, increasing the burden to review them and discern which were created with actual human thought and care. Like all technologies, AI presents a mixed bag of benefits and costs.

Use Case 2: First-Draft Evaluation Planning

AI could also revolutionize how evaluators brainstorm in the design phase of an evaluation. Evaluation planning is both critical and time-consuming, involving developing rationales, selecting an evaluation approach, developing a framework such as a theory of change or logic model, generating criteria, selecting indicators, etc. ChatGPT can contribute to many of these processes. In testing ChatGPT for the preparation of this article, we were able to produce a set of criteria and even a logic model for evaluation of a homeless shelter. What resulted were not finished products, but a stunningly good set of starting points, the only input being a few keystrokes.

Our conviction is that thoughtful evaluative engagement will be needed as much (if not more) in the future as in the past—to generate, review, confirm and finalize all aspects of a design. However, our tests indicate that utilizing ChatGPT can serve as a helpful starting point that could increase the quality of a design by presenting ideas that might not have surfaced within a traditional process.

Other Possible Uses

With these examples in mind, the number of potential applications is vast. For example, AI could draft survey questions or qualitative protocols—serving as a "fantastic sparring partner for thinking" as noted by Silva Ferretti at a recent virtual symposium on AI in evaluation (Simon, 2023, para. 6). By quickly synthesizing vast amounts of existing text data, AI could also be useful for research on evaluation by providing a quick snapshot of most common approaches or understandings. There are also future possibilities related to data analysis. For instance, as of this writing Microsoft has announced that they are developing a virtual assistant known as Copilot, which will soon be embedded in Office products (Microsoft, 2023). This will allow, with simple text prompts, a worker to instruct an AI program to analyze an Excel spreadsheet with little input. With properly structured data, this seems likely to greatly speed the process of identifying trends, outliers, and patterns. Microsoft Copilot will also have the capability to take a Word document and convert it to a PowerPoint presentation, substantially

speeding the process of generating multiple report formats.

As OpenAI rolls out its APIs to third-party vendors (Kan, 2023), ChatGPT functionality will also be embedded inside other applications. Text queries could be available within MAXQDA or similar qualitative analysis software to speed qualitative coding and analysis. Momentum toward these applications is building quickly. There is a potentially extensive list of benefits from AI—but these benefits also come with new costs, risks, and challenges to evaluation practice.

Potential Challenges of AI for Evaluators

Based on our understanding of the history of advancing technology, one thing seems clear: AI is likely to become ubiquitous. The promise is that these advanced tools will feel like ultra-intelligent personal assistants helping with things we simply do not want to do, or cannot do alone. While we summarized the potential benefits for evaluators above, this section will outline the expected challenges for the field of evaluation.

Challenge 1: Output, Output, and More Output

We see the cheapening of information products as a potential risk to the field. A technology that can generate endless reports will very likely change the nature of the report itself. Will a report cease to be meaningful?

With the ability to feed data into an AI tool that generates a report, the cost of evaluation will be much lower. AI-generated reports may indeed produce some useful feedback to inform program managers—especially for those with few resources available for more in-depth evaluation. However, this may also contribute to new problems for evaluators and program managers.

Good evaluations have always needed to be constructed and conducted with care. An evaluation carried out without regard to theory or practice, without considering the underlying evaluative logic or criteria being used, could prove to be useless, misleading, or even harmful. At least in the short term, for instance, it is unlikely that AI will be able to flag serious shortcomings or biases in an evaluation design. This automation could result in the hollowing out of the evaluative process, which would represent a great loss for programs as well. Patton (2011) argues that the "process use" of an evaluation has cascading benefits for an organization over and above the simple findings of the evaluation (p. 142). If the creation of an evaluation has little human involvement, this valuable process use would be entirely eroded.

On the other hand, thoughtful evaluators could leverage the new ease of producing evaluation outputs to greater value for stakeholders. The AI tools could shift the balance of contributions evaluators provide—reducing the time spent producing outputs and increasing time devoted to formative and utilization-focused services. In this scenario, evaluators could work alongside organizations to create a system that streamlines developmental processes and focuses the limited human energy on process use. Once a solid system is put in place with the guidance of a trained evaluator, then the organization could generate reports as needed. While AI tools might lower the value of evaluation products, as Patton highlighted, much of the utilization value has always been in the process—and these tools could provide a practical method to shift energies toward engaging and participatory processes.

Challenge 2: Market Disruption and Emergence of New Roles

As the barrier to entry for evaluation products becomes lower, will demand for evaluators go down? Will the big evaluation firms who are most able to leverage AI technologies simply expand their dominance? If AI results in the average evaluation product being of lower quality and utility, is society likely to discredit evaluation as a whole? While these outcomes are conceivable, they are by no means inevitable. Consider the famous anecdote of what happened with banking after the rise of the ATMs: The technology actually (in the first thirty years of its existence, at least) led to more teller jobs (Pethokoukis, 2016).

These possibilities bid us to consider the alternative paths the future might take. For instance, it is possible that evaluations will simply become cheaper—and therefore more accessible to more organizations. Could this launch a golden era of evaluation for small nonprofits, for instance? Or maybe an expanded use of evaluation by for-profit businesses or social ventures? Or perhaps it will empower the development of much larger, information-rich evaluation processes for those who can afford them.

Challenge 3: The Alignment Problem

We have already mentioned the challenge of the inscrutability of AI design—a black box into which humans are not able to peer. This has concerning implications for AI's influence over human decisions and related real-world consequences. Without a practical means to understand the details of AI design, humans will have little ability to nudge the process in ways that we feel are consistent with our values—such as equity. This quandary has enormous implications for what is famously known as the alignment problem. That is, if we cannot be sure exactly how a system is operating, we can't be sure that it is operating in accordance with our values. Equally as complex: How do we even begin to arrive at a set of agreed-upon values at such a large scale for a system that will be operating in so many contexts, across so many situations, and in so many cultures? As Gabriel (2020) starkly puts it:

> How are we to decide which principles or objectives to encode in AI—and who has the right to make these decisions—given that we live in a pluralistic world that is full of competing conceptions of value? Is there a way to think about AI value alignment that avoids a situation in which some people simply impose their views on others? (p. 412)

Stunningly, this central question about AI and value alignment is at its heart an evaluative question, and one that our field is uniquely positioned to answer. Perhaps this is not surprising, considering Scriven's (2013) assertion that evaluation is a fundamental cognitive activity, present in our own species for the past 3.5 million years, and also visible in others. It seems natural, then, that we should expect to see these challenges arise when we build a machine that is meant to mimic, and indeed surpass, the human intellect.

Practically speaking, this has real implications for evaluators. In the case of an AI that has not achieved the level of superintelligence (surpassing the abilities of a normal human) this already starts to be difficult. Consider the case in which I have an AI tool analyze some survey data, looking for trends, and perhaps even flagging some "unfavorable" outcomes. It would be difficult to know if values implicit in the AI that were used to flag these patterns are unassailable, or to even contend with them in the first place. This problem seems to compound when AI does rise to the level that surpasses our own abilities. One can see a scenario in which AIs generate reports by which programs and agencies (and therefore real people) live and die. Will evaluators be able to develop approaches to understand these systems so that we can confirm that our values are reflected?

Challenge 4: Using Our Agency

Given all of the challenges, why should evaluators accept AI technologies? Why should we not resist? There are significant concerns about the ethics and safety of AI. However, the wheels of industry, finance, and Silicon Valley are already turning swiftly. Like it or not, in our estimation, this is a force that cannot be stopped. If our field (like every other field) is to survive the transformations that are coming for all types of knowledge work, we will need to be prepared for change. We must adapt or become irrelevant—just as modern evaluators would be if they were still using typewriters, slide rules, and carbon paper.

In a recent interview on AI, economist Bryan Caplan stated, "All progress is bad for somebody. Vaccines are bad for funeral homes. The general rule is that anything that increases human production is good for human living standards" (Cantor, 2023, para. 10). While this may be true on average, it also ignores the fact that new technologies introduce new forms of inequality and human suffering. We can and should be proactive, creating new ethical standards for society in general, but also for evaluators specifically. While the change is inevitable, and there is reason to be optimistic, it does not mean we can let down our guard and take our hands off the wheel. We must use what agency we have to chart a course toward just and equitable AI implementation.

Conclusion

In this article we have addressed some possible use cases and challenges presented by AI in evaluation practice. Our world is about to change dramatically—possibly as much as or even more than with the introduction of the internet, smartphones, and social media. With all of this in view, what can evaluators do now to prepare for this future?

First, evaluators should start getting to know the tools. Play with ChatGPT. Play with Google's Bard and Microsoft's Copilot. Think about the applications they might have on your work, and do your own tests. It is likely that the first to adopt these tools will find particular rewards. The work of exploring AI should also include significant research on evaluation efforts, as this will allow researchers to establish in empirical terms the

effects that integrating this new technology is having on the field.

It is also necessary that evaluators be transparent and proactive in their use of this technology. Our stakeholders must be notified about how we are leveraging the technology and where our human expertise is harnessed—lest they get the sense that we are obsolete or, worse, carelessly wielding a powerful technology. Along those lines, the American Evaluation Association should consider updating the guiding principles and the evaluator competencies to include guidelines for ethical AI use.

Third, evaluators should think critically about how AI will affect our work. We should not simply think in binary terms of better/worse. The implications are likely to be far more nuanced, and we will need to be less reactive and more thoughtful, creative, and adaptive.

Lastly, we can take a deep breath and calm down. All of society is living through this moment together. Evaluators are not uniquely affected by all of this, and there is still so much unknown. It is likely that a year on from this publication, the situation will be much clearer. We are optimistic that we can work together to chart a course for positive change.

References

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). *Large language models in machine translation*. Google, Inc. http://research.google/pubs/pub33278.pdf

Bruce, K. (2023, March 5). *ChatGPT and evaluation: 3 key takeaways* [Post]. LinkedIn. https://www.linkedin.com/pulse/chatgpt-evaluation-3-key-takeaways-kerry-bruce/

Cantor, M. (2023, April 7). This economist won every bet he made on the future. Then he tested ChatGPT. *The Guardian*. https://www.theguardian.com/technology/2023/apr/06/chatgpt-ai-bryan-caplan-interview

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20–23.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335.

Gabriel, I. (2020) Artificial intelligence, values, and alignment. *Minds and Machines, 30*, 411–437. https://doi.org/10.1007/s11023-020-09539-2

High-Level Expert Group on Artificial Intelligence. (2019). *A definition of AI: Main capabilities and disciplines*. European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

Kan, M. (2023, March 2). ChatGPT is coming to an app near you: OpenAI launches API for its chatbot. *PC Magazine*. https://www.pcmag.com/news/chatgpt-is-coming-to-an-app-near-you-openai-launches-api-for-its-chatbot

Marcus, G. (2022, December 1). How come GPT can seem so brilliant one moment and so breathtakingly dumb the next? *The Road to AI We Can Trust*. https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant?

Microsoft. (2023, March 16). Introducing Microsoft 365 Copilot—A whole new way to work. *Microsoft 365 Blog*. https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/

OpenAI. (2022a, September 28). DALL·E now available without waitlist. *OpenAI Blog*. https://openai.com/blog/dall-e-now-available-without-waitlist/

OpenAI. (2022b, November 30). Introducing ChatGPT. *OpenAI Blog*. https://openai.com/blog/chatgpt/

Patton, M. Q. (2011). *Essentials of utilization-focused evaluation*. SAGE Publications.

Pethokoukis, J. (2016, June 6). What the story of ATMs and bank tellers reveals about the 'rise of the robots' and jobs. *AEIdeas*. https://www.aei.org/economics/what-atms-bank-tellers-rise-robots-and-jobs/

Scriven, M. (2013). The foundation and future of evaluation. In S. I. Donaldson (Ed.), *The future of evaluation in society: A tribute to Michael Scriven* (pp. 11–44). Information Age Publishing.

Simon, N. (2023, April 15). ChatGPT in evaluation—An opportunity for greater creativity? *University World News*. https://www.universityworldnews.com/post.php?story=20230412111133714

Walker, L. (2023, March 28). Belgian man dies of suicide following exchanges with chatbot. *The Brussels Times*. https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt/