Theories and Methodologies

A(I) University in Ruins: What Remains in a World with Large Language Models?

KATHERINE ELKINS

The task of my piece, as the last word, is to reassure you that, despite all the bad news, our future will be okay. So here goes.

Everything will be okay. Also, our field is more important than ever.

I believe both these statements to be true, but not for the reasons you might imagine.

You may think I'm about to tell you that we'll be just fine because the recent AI craze is nothing but a hype bubble. For every headline that heralds a breakthrough, another announces some fundamental flaw.

Takedowns of AI can be reassuring, and they are both a symptom of and a salve for our shared anxiety. But they can also mislead. Thanks to the well-known Moravec's paradox, AI is often quite bad at things that are easy for us, while—paradoxically—quite good at things that are difficult. Finding an example of an AI weakness and generalizing by analogy to our human capabilities can be wildly inaccurate.

Another reason it's difficult to generalize from weaknesses is that critiques often rely on an outdated model, like *ChatGPT* 3.5, and fail to take into account the rapid rate of change. Or, rather than offer insights into the model's true capabilities, they reveal more about the value alignment process, which can sometimes distort a model in problematic ways.[1]

What might surprise you is that, for those of us working with large language models (LLMs) since their earliest inception, advances in AI have been gradual and continuous. Limitations in one model are overcome by the next, and progress comes step-by-step.

KATHERINE ELKINS is professor of comparative literature and humanities at Kenyon College, where she codirects the KDH Lab. She is the author of *The Shapes of Stories* (Cambridge UP, 2022), a contributing editor of *Philosophical Approaches to Proust's* In Search of Lost Time (Oxford UP, 2022), and a principle investigator for the US AI Safety Institute and the Notre Dame / IBM Tech Ethics Lab. Her current research explores storytelling, literary translation, and AI ethics.

PMLA 139.3 (2024), doi:10.1632/S0030812924000543

In Franz Kafka's parable "Before the Law," a man stands before a fearsome gatekeeper and is told there are many more gatekeepers, each more powerful than the last. So, too, there are many more AI models to come, each more powerful than the last. In the parable, the man standing before the gate realizes only too late that the gate is meant for him. Reader, this gate is meant for us, and we should step through before it closes.

Now it's true that we AI researchers disagree among ourselves about just how quickly current limitations will be overcome, and by what means.[2] While AIs are certainly not on a par with our most creative and original thinkers, we are well on our way to an AI that can automate more typical levels of human creativity and intellect.[3] As Timothy Laquintano and Annette Vee write, many are already using LLMs for everyday writing. Seth Perlow aptly suggests that writers' reactions can seem, well, reactionary.

How, then, is this good news for our profession?

For one thing, we have many new tools for exploring long-standing questions about things like the meaning of language, as Aarthi Vadde does, the shapes of stories (Elkins, *Shapes*), the role of the author (Elkins, "AI"), or the task of the translator (Elkins, "In Search"). While human-centered AI teaching and research (Chun and Elkins, "Crisis") used to be accessible only to those fluent in artificial languages, programming in natural languages will democratize new ways to answer questions that matter to us.

Also, this conversation is taking place within forums of the Modern Language Association, and these are language models. Major breakthroughs in artificial intelligence came not from manipulating symbols and logic but from surfacing patterns in vast quantities of text. This is a skill those of us working on long, complex novels know quite well, and it should comfort us that this—more than any abstract symbol manipulation—has proven a key to intelligence.

Just how much of the world is contained in the word is now debated by AI researchers.[4] Our world has become theirs, and I've had model developers defer to me in conversations, insisting, "You tell me: you're the language expert." Indeed, we are, and it's our field as much as theirs.

Working with LLMs is not just playful experimentation or fun with language games. These models are black boxes, and while we understand the basic mechanisms, we cannot understand exactly why or how they produce a particular output. It's a bad idea to generalize about a model from a single output or two, but we can begin to get a sense of it from more extensive auditing of outputs, and this analysis requires the kind of careful qualitative work of close reading that we excel at.

Matthew Kirschenbaum and Rita Raley emphasize this kind of qualitative assessment, so let me say more about why it's so important and what we can bring to the conversation. A traditional computer science approach subjects GenAI models to batteries of tests in order to establish quantitative benchmarks. But these fail to take into account the kind of qualitative differences that are so important for establishing model capabilities. This is why I advocated for the MLA to join the US AI Safety Institute as an inaugural consortium member. Our MLA AI research team will be assessing model capabilities and, alongside other Safety Institute members, making recommendations on safety and policy based on our qualitative assessments.

To give a concrete example of what this looks like, last summer I collaborated on an ethical audit of leading chatbots (Chun and Elkins, "Informed AI Regulation"). We evaluated a wide variety of ethically fraught situations, asking the LLMs to reason through ethical problems by carefully weighing pros and cons and explaining the most ethical course of action in situations in which there is no easy answer. The quantitative measurements were important for benchmarking performance, especially when it came to assessing troubling authoritarian tendencies in some models. But qualitative differences were even more revealing, and these qualitative differences could have real-world consequences should these models steer AI autonomous agents or, even more frighteningly, power autonomous weapons.

I've presented this research to fellow AI researchers, and it's been very well received by experts in the field.[5] But it's been less well received by fellow humanists. One anonymous peer reviewer found it laughable that we could even talk about reasoning in an LLM. And yet, if you search for articles on AI reasoning, you will see that it's actually a robust field of AI research.[6]

Don't get me wrong. As with so much in the field, researchers debate among ourselves what we mean by words like "reason" and "intelligence." These are legitimate debates that depend on careful experimentation with models. But to participate in these language debates without reference to the details is like making sweeping judgments about a novel without reference to the text.

Of course, you could accuse me of anthropomorphizing AI. To be clear, I don't believe AIs are conscious or sentient, and it's unclear whether they ever will be. But many AI researchers, including Geoffrey Hinton, known as the "godfather of AI," make strong claims for AI "understanding." As Perlow points out, the world of AI researchers is now focused on the kind of linguistic ambiguity that we know all too well.

When I asked a developer of one of the leading LLMs what he thought about all the hand wringing over using terms once reserved for humans, he replied, "Gatekeeping!" I'm certainly not advocating for playing fast and loose with language—far from it. Rather, Junting Huang is right to invoke Ludwig Wittgenstein's language games. Terms now commonplace in AI research do not presuppose a perfect correspondence between humans and machines but a situation in which, as several of the essays point out, the line between the two is blurred.

Because of the black box nature of these models, there is actually a new field employing methods from psychology to assess these LLMs. Many researchers are testing them for personality traits (Hilliard et al.), while others have probed the models for theory of mind (Xu et al.). The analogy is imperfect, but the kind of qualitative close reading and analysis we performed for our ethical audit is not entirely dissimilar from trying to discern an author behind a corpus of texts. What aspects of an LLM's output reflect the voice of the prompt we give it, merely adopting a tone or style in response to our commands like an author might voice a character while holding entirely different views? And what outputs give us a sense of what the LLM may really be like, of how it might behave were it to lose its programmed guardrails and act as an autonomous agent, or of what it might do were it to be commandeered by a bad agent?[7]

These assessments also include the important work of surfacing and evaluating bias, work that my lab has been doing for several years. Only by assessing all these downstream linguistic outputs can we determine the bias or fairness of the model. This work is pressing: biased AI has been in our world longer than LLMs, and AI bias already harms many. The methods for debiasing LLMs are far more complex than the debiasing process for "white-box" AI models, however, and require new methods and interventions (Elkins, "Reading").

What's more, all AI models will be biased by their training data. Many leading models show a strong bias for Western values. They can also exhibit new biases created by the attempt to align their outputs using human feedback. Whose values do we choose in performing such alignment? Human-AI alignment will ultimately rely on human-human alignment, an alignment that will—and should—always be imperfect. Who would want to flatten the many different ways of seeing the world that are evidenced in the linguistic and cultural variety that we teach and study?

For this reason, our profession can have an impact in advocating for linguistic and cultural diversity in the models we create, as Eduardo Ledesma so eloquently argues. Instead of trying to create an ideal "objective" model, the solution may be to create a wide range of models trained on different languages and cultural datasets. One of the best ways to do this is by advocating for open source models, which can be fine-tuned to represent different cultures. Advocating for open-source models also means advocating for more democratized access to models in the Global South, for instance, as opposed to concentrating proprietary models in a few Western countries.

We probably don't want to live in a world in which a few American tech companies control a handful of models that are predominantly anglophone and have a Western bias. But whether and how open-source models should be regulated is a complex issue, since such freely available models could also be used by bad actors. Many of us are involved in research focused on how use to open-source models responsibly, and we need more researchers working on this.[8]

To be clear, I'm not advocating for a monolithic or single approach to engaging with AI. In this, I follow Kirschenbaum and Raley's enumeration of several quite different paths. It would be easy to fall into arguments about the priority of one avenue over another, but this distracts from the work at hand. For those of you asking, "Who am I to weigh in on any of these issues?" I would ask you whether you believe these questions are best left to a handful of AI experts, many with quite narrow specializations. Perhaps the better question is, Who if *not* you?

But if this is just too "applied" humanities, too in the weeds, might I recommend that we expand our field of *theory* to include LLMs as well? AI researchers are now asking questions like whether words are the same thing as actions, and whether thought requires language. We argue over whether humans will be happy without work and whether there are more good actors than bad in the world.[9] We debate whether our new technology will give us superpowers by augmenting existing capabilities or foster intellectual and creative atrophy as we grow increasingly reliant on our tools. Again, I ask, are we content to let these conversations happen without us? Who if *not* us?

Of course, to engage in any of these activities is a big ask, and you may be thinking that this is not the job you signed up for when you entered the profession. While it's easy to criticize AI from afar, it takes quite an investment of time to come up to speed and stay current in the field, which is developing at breakneck speed. I don't want to minimize this challenge, since it's difficult for even those of us established in the field. What's more, I second Meredith Martin's claim that much of this work is not adequately recognized in the academy. This goes not just for the intensive reworking of classes and curricula but for participation in collaborative communities as well as research that doesn't take the form of a peer-reviewed monograph.

This kind of work will not appeal to everyone, even when rewarded, and I turn now to those for whom neither the applied nor the theoretical avenues prove tempting. But first, let me do a bit of speculative design, laying out the various forking paths that may or may not await us in a future with AI.

What the slightly longer term holds depends on whom you ask. In the most traditional view, economists and historians insist that the past predicts our future. If history serves, concerns over automation and job loss are inevitably followed by new growth and job creation (Ekelund). This time will be like last time, so we need only await the new jobs that we can't yet imagine. We can educate our students for a new-old future, a future that we will evolve to meet, just as we have evolved in response to earlier technologies. Our universities will morph and change like the turn of a kaleidoscope, the same colors and shapes reorganized into new patterns. The university of old will be in ruins, true, but a new university not too dissimilar will rise from the rubble.

Unfortunately, most who work in AI don't share this view, believing that intellectual and creative automation pose a challenge quite different from earlier technological disruptions. Among forecasters, we can choose between the utopian and dystopian visions. Techno-optimists like Marc Andreesen and Elon Musk write of a postlabor world with an infinite supply of goods. Our future will be similar to the utopias many of us have taught as fictions. The university will be in ruins, but it should not concern us terribly. The ruins will be replaced by a world of plenty, and we will all be freed from the labor of education, enjoying instead the fruits of AI labor. If this utopia awaits, then we might best prepare by organizing our reading groups with our utopia-island reading lists. We might also do well to glean lessons from our utopian fictions as we consider what human flourishing means in a workless world.

On the other side, the most extreme version of the dystopian vision is held by the so-called AI doomers, who believe that AI carries the threat of human extinction. This is not a popular view, and most of us don't see this as an immediate risk, but I for one am glad that there are people worrying about this and creating tests to screen for dangerous capabilities (Phuong et al.).

The less extreme dystopian vision holds that AI labor will not create a land of plenty. Instead, we will find ourselves in a world of even greater inequality than our present day. A very few will still hold jobs, but a larger and larger number will be underemployed or even unemployed. This is yet another kaleidoscopic version of our existing world of growing inequality, only amplified. For this world, we could do worse than rely on our field's robust tradition of critique of capitalism to help guide us.

To return to the question of what this means for our profession, let me just spell out the difficulty we face. Since the release of *ChatGPT*, one of the most surprising things about generative AI is that it threatens to automate intellectual and creative tasks. It won't be immediate, but there's no question that AI research, writing, and translation will automate some of what we have typically thought of as our domain. While some of us engage more directly with AI, others had best start imagining what the rise of AI means for our everyday practice of teaching language and literature.

There's actually some good news here. In either postlabor outcome, we may find ourselves liberated of the vocational imperative. We would no longer need to worry about whether we prepare our students for jobs beyond the university, nor would we need to worry about how much our students conform to the desires of employers. We may even be freed from having to insist that the humanities offer what employers most want: graduates who can think creatively and critically and who can write well. It could be a return to the "uselessness" of the humanities, the kind of "purposeless thinking" that Martin describes. For some of us, this would be a positive development, since it signals a vocation that lies outside the purview of the neocapitalist imperative for utility.

For the first time, moreover, we're hardly alone in our crisis. In a world of intellectual and creative automation, many white-collar jobs will be at risk of elimination, and these are jobs we've typically relied on to justify the value of higher education. While not yet perfect, generative AI is increasingly good at coding and data analytics, which means our STEM colleagues may soon share our crisis, and they have farther to fall, having built out to meet huge demand. Affected, too, will be creative writing and other fine arts faculty members, our colleagues who prepare our creative class. We are now, finally, all in it together.

When utility is stripped away, what remains for higher education? In fact, the humanities may fare pretty well. Reading philosophy and poetry, plays and novels—connecting to the many and varied experiences of being human past, present, and (speculative) future: it's hard to imagine a better way to spend our human-centered higher education time.

Admittedly, I haven't had time to do much of this lately. Like many of the other writers in this series of essays, I've been pulled into discussions on topics ranging from AI regulation to the future of work. What keeps me up at night are not questions about how to identify students who've cheated using AI. What I lose sleep over are the bigger questions about what the future holds and how we can best prepare. These include the larger social issues I outlined, the challenges to our profession and to higher education more generally, and current technical challenges that, I firmly believe, should be solved with our help. How can we ensure cooperation in a world of multiagent AI networks? How might we control an AI that is more intelligent than we are? These are just some of the challenges many of us are working on.

Even though I spend my days working on these issues, this response is not a judgment on those who choose not to. There are more than enough paths to prepare for, and I would like to believe that, once we've solved some of these alignment problems—as I hope we will—I can return to my old friends, the authors who line my bookshelves and await quieter and more contemplative days. While some of us engage with AI more directly, others will need to continue to fight for the value of learning languages and the importance of reading literature.

I'm not arguing that our work should serve AI or even, for that matter, the university. Instead, I'm advocating for an AI that would serve us, that would answer the questions that we determine are the most important and align with the values we choose, however impossible that task may seem. I am asking if we're willing not just to ask the big questions but to decide what new questions need to be asked. And I'm suggesting that we embrace a more influential role in shaping a future with AI, even if that work takes us away from our more traditional practices.

Some of us will need to keep those more traditional practices in place for a time when these issues have become—as I hope they will—less pressing. But if you feel inspired to join in this AI work, let me just say, you are needed. The gate is for you, and we will tackle the more ferocious gatekeepers together as we meet them, one by one.

Notes

1. This is likely the case for Google's *Gemini*, which during the initial rollout depicted American founding fathers and Nazis as Black (Allyn).

2. One aspect that surprised even AI researchers is that scaling (i.e., building larger and larger models) continues to yield advances. Still, most agree that we need new techniques to produce the next breakthrough. Recent work has focused on building an ensemble of smaller expert models much like the connected but distinct regions of a human brain and adding additional types of nonlinguistic training data to augment knowledge of the world.

3. Some say we are still a long way off—for example, Yann Lecun (New York University and Meta) and Christopher Manning (Stanford University). Others, like Sam Altman (OpenAI) and Elon Musk (Grok), suggest breakthroughs could happen in the next five years, if not sooner. The general consensus among thousands of AI researchers surveyed in 2023 was that there is at least a fifty percent chance of a major advance by 2028 (Grace et al.). Forecasters at Google's DeepMind study see a likelihood of dangerous capabilities by 2029 (Phuong et al.).

4. Many AI researchers agree that words contain a surprising amount of knowledge about the world. See a discussion of this phenomenon in Altman's recent interview with Lex Fridman ("Sam Altman"). For a (somewhat) contrary opinion, see Fridman's interview with Lecun ("Yann Lecun").

5. Results were shared with members of the Open Innovation AI Research Community.

6. For example, see Zelikman et al. The paper's references give a good sense of just how established the field is. There is even research into "self-reasoning" to assess dangerous capabilities (Phuong et al.).

7. Of all the dangerous capabilities that DeepMind researchers assessed, including cybersecurity threats, self-proliferation, and self-reasoning, the most advanced current capability lies in the realm of persuasion and deception (Phuong et al.).

8. One such group is the Open Innovation AI Research Community, which has key members from the Global South. A positive impact of AI has been the growing localization of content and a rise in multilingual web pages ("Usage Statistics"). This linguistic diversity is at the forefront of conversations in our Machine Translation community, most recently at the March 2024 SlatorCon (Txabarriaga). Unfortunately, as Ledesma notes, the MLA AI research community lags behind in this regard.

9. This latter question concerns the open-sourcing of models. The argument goes that if most people are good, open-source models in the hands of more good actors could empower us to combat misuse by bad actors.

Works Cited

Allyn, Bobby. "Google Races to Find a Solution after AI Generator Gemini Misses the Mark." *NPR*, 18 Mar. 2024, www.npr.org/2024/03/18/1239107313/google-races-to-find-a-solution-after-ai-generator-gemini-misses-the-mark.

Chun, Jon, and Katherine Elkins. "The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centred AI." *International Journal of Humanities and Arts Computing*, vol. 17, no. 2, 2023, pp. 147–67.

———. "Informed AI Regulation: Comparing the Ethical Frameworks of Leading LLM Chatbots Using an Ethics-Based Audit to Assess Moral Reasoning and Normative Values." *ArXiv.org*, 9 Jan. 2024, arxiv.org/abs/2402.01651.

Ekelund, Henrik. "Why There Will Be Plenty of Jobs in the Future—Even with AI." *World Economic Forum*, 26 Feb. 2024, www.weforum.org/agenda/2024/02/artificial-intelligence-ai-jobs-future/.

Elkins, Katherine. "AI Comes for the Author." *Poetics Today*, vol. 45, no. 2, 2024, pp. 267–74.

———. "In Search of a Translator: Using AI to Evaluate What's Lost in Translation." *Frontiers in Computer Science*, vol. 6, 12 Aug. 2024, https://doi.org/10.3389/fcomp.2024.1444021.

———. "Reading GenAI: The Trouble with Bias." Reading Generative AI: Theory, Data, Critique, Matthew Kirschenbaum and Rita Raley, presiders. MLA Annual Convention, 5 Jan. 2024, Philadelphia.

———. *The Shapes of Stories*. Cambridge UP, 2022.

Fairman, Gabriel. "Translating Worlds: AI Expert's Adventure from Pages to Pixels (with Kate Elkins)." *Merging Minds*, 2024. *Apple Podcasts*, podcasts.apple.com/us/podcast/translating-worlds-ai-experts-adventure-from-pages/id1727352682?i=1000648935735.

Grace, Katja, et al. “Thousands of AI Authors on the Future of AI.” *ArXiv.org*, 5 Jan. 2024, arxiv.org/abs/2401.02843.

Hilliard, Airlie, et al. “Eliciting Personality Traits in Large Language Models.” *ArXiv.org*, 15 Feb. 2024, arxiv.org/abs/2402.08341.

Hinton, Geoffrey. “Will Digital Intelligence Replace Biological Intelligence?” The Romanes Lecture, University of Oxford, 19 Feb. 2024. *YouTube*, www.youtube.com/watch?v=N1TEjTeQeg0.

Kafka, Franz. “Before the Law.” *A Hunger Artist and Other Stories*, translated by Joyce Crick, Oxford UP, 2012, pp. 20–22.

Phuong, Mary, et al. “Evaluating Frontier Models for Dangerous Capabilities.” *ArXiv.org*, 20 Mar. 2024, arxiv.org/abs/2403.13793.

“Sam Altman: OpenAI, GPT-5, Sora, Board Saga, Elon Musk, Ilya, Power and AGI.” *Lex Fridman Podcast*, episode 419. *YouTube*, www.youtube.com/watch?v=jvqFAi7vkBc. Accessed 21 Mar. 2024.

Txabarriaga, Rocío. “Should AI Be Trusted as It Gets Better at Translation?” *Slator*, 28 Mar. 2024, slator.com/should-ai-be-trusted-as-it-gets-better-at-translation/.

“Usage Statistics and Market Share of Content Languages for Websites, February 2020.” *W3Techs*, 2020, w3techs.com/technologies/overview/content_language.

Xu, Hainiu, et al. “OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models.” *ArXiv.org*, 14 Feb. 2024, arxiv.org/abs/2402.06044.

“Yann Lecun: Meta AI, Open Source, Limits of LLMs, AGI and the Future of AI.” *Lex Fridman Podcast*, episode 416. *YouTube*, www.youtube.com/watch?v=5t1vTLU7s40. Accessed 21 Mar. 2024.

Zelikman, Eric, et al. “Quiet-STaR: Language Models Can Teach Themselves to Think before Speaking.” *ArXiv.org*, 18 Mar. 2024, arxiv.org/abs/2403.09629.