Lecture | December 2024
Ricardo Galliano Court, Assistant Dean for Academic Integrity and Undergraduate Research
Weinberg College of Arts and Sciences, Northwestern University

Illusions of Understanding

The hype and promise of generative artificial intelligence run the range from the mundane to curing cancer, solving climate change, and transforming the global economy. Closer to reality, AI promises productivity, efficiency, and objectivity. For students and researchers alike, AI promises to automate tasks that are seen as drudgery, opening up possibilities for more rewarding activities. AI however, poses risks to academic integrity and understanding. This seems like a truism, but the risks are far from obvious.

In my work as Dean for Academic Integrity in Weinberg College, I have spoken with colleagues who are AI's detractors and its boosters and very few people in between. Personally, I remain a skeptic who nevertheless sees the possibilities of examining massive texts, data sets, and contexts in ways that we can barely envision. While my interests lie mainly with undergraduate education, the point of that education is to train students to be or at least to think like scientists. As researchers, teachers, and students we all should consider the ethical concerns over the use of AI.

Today I would like to leave aside worries that generative AI will allow students to claim grades that they do not deserve and that we will no longer be able to rank students by merit represented by those grades. These are valid concerns and yet, they are mere proxies of what, in my view, should concern us.

How are students actually using generative Ai? When we interview and poll undergraduates on their uses of large language models they respond with interesting answers. Students say they employ AI to:

"research"
"brainstorm"
Outline
Edit
Summarize and to generate comments

Despite the undeniable power of the models, there are hidden issues with all of these elements.

- While students see "research," what they get is information generated by Large Language Models that reads as plausible natural human language which is not necessarily true.
- When AI "brainstorms," the ideas that it generates are the composites of analyses and critiques that have already been done, complete with their errors, bias, and misconceptions built in. Less "brainstorm" and more patchy clouds and drizzle.
- GenAI's outlines conform to the most common framing of a given problem, and thus are strongly resistant to novel perspectives.
- AI has entered (or invaded, infiltrated) most editing software—students can be unaware when AI is doing more than correcting their syntax and grammar. The AI is manipulating their words, and perhaps their ideas.

- Finally, AI summaries of concepts show evidence of the same misconceptions and misunderstandings as the above.

The reality of what AI can and can't do should give all students pause before they use it. But the problem as I see it goes deeper – and started before GenAI came on the scene.

I call it "The Information Management Problem."

According to students' understanding, *The Information*, as they call it, is out there on the web, and their task is to format *The Information* according to the requirements of their professor. Students' default is not to apply evidence in making a novel argument; rather it is to repeat what is already known. This way of thinking about the project of research can lead pretty easily to plagiarism, and it is easy to see how both the attitude and the risk of potential plagiarism might be turbocharged by AI. The web created the illusion that anything knowable was just a web-search away. AI has the potential to narrow that view even further: any question that can be asked is given an instantaneous and reasonable sounding answer. I tell my students that we do our best academic writing on questions for which we don't yet have answers. When every question instantly receives a 'plausible' answer what we lose is discovery.

When discussing the potential and the risks of AI, it is difficult to separate probabilities from the vaporware and doomsaying, sometimes presented by the same people. I want to narrow my comments today, as I have in conversation with undergraduates, to how we accurately represent our work in good faith. In addition to built-in bias, high-energy consumption, false and misleading imagery, bots, misleadingly-termed "hallucinations", the degradation of AI trained on AI, and the alignment problem, Messeri and Crockett, break down four ways that scientists are beginning to use AI, all of which have significant flaws and ethical ramifications:

AI as:

As "Oracle", to objectively and efficiently search, evaluate, and summarize literature.

As "Surrogate", to generate plausible data points, in place of data too difficult to collect.

As "Quant", to analyze vast texts or data sets, impossible or too time consuming for humans.

As "Arbiter", to objectively evaluate studies, with the hope of removing human bias.

Messeri and Crockett intersect these four, saying "the large datasets that Surrogates produce require Quants to analyze them. Furthermore, the efficiency promised by Surrogates and Quants will yield even more publications for Arbiters to adjudicate and for Oracles to digest and summarize. These visions also reinforce one another by converging on two broad goals: to enhance scientific productivity by overcoming scientists' limited time, attention and cognitive capacities; and to enhance scientific objectivity by overcoming scientists' subjectivity and bias. Achieving these goals is expected to improve scientific understanding."

AI boosters are promoting their platforms as "trusted partners" and extensions of users' minds rather than mere tools, without an awareness of the risks involved. For example, Researchers run the risk of inadvertently restricting their research to studies conducive to AI assistance and its promise of productivity, efficiency, and objectivity. The analog to undergraduates restricting *The Information* to that which is readily served up to them on their devices should be obvious and the same pitfalls apply.

A powerful way to look at plagiarism is as a failure to engage an ongoing scholarly conversation in good faith, and instead, creating the illusion of doing so. Whether one does this knowingly or inadvertently makes little difference because the product of research can lead our readers astray. Failure to recognize the limits of AI tools risks creating a similar bad faith illusions in research. It is unclear to what degree are we and all the scholarship we consult narrowing our perspectives to those made convenient by our tools while presenting our work as comprehensive. The risk grows in parallel to AI's capacities. The one guiding principle I tell my students is: *Don't pretend to do something that you are not doing.*

AI promises to ameliorate our cognitive limitations and abet our cognitive miserliness (a wonderful term coined by a pair of Bangladeshi researchers). The draw of AI ranges from allowing us to do what we are otherwise unable do to, to simply avoiding what we would rather not do. Throughout that range, AI will produce results the veracity of which we are either unable or unwilling to confirm. Students who use generative AI so that they can prioritize other activities think that they have their cake and are eating it too; in truth, they cannot know what they do not – in truth, they can't know the ingredients that went into the cake, or indeed, if the cake is actually cake or just foam rubber. Researchers who over rely on AI, even when the results are accurate, are at risk of the same over-assurance. Messeri and Crockett imagine a near-future where, "we produce more but understand less."

The illusion of "Prompt Engineering"

The output of generative AI has been designed to sound convincing. The students with whom I have talked about generative AI are all skeptical of the outputs, their linguistic, socioeconomic, cultural, and racial bias, in addition to issues of accuracy. At the same time, they find it nearly impossible to avoid the allure of the plausible and conversational tone of the output, for example, the chatbots' engineered insistence on being a moral agent, passing judgments on queries that are overly risque. Students tell me that they are disappointed by the typical outputs of generative AI, but they envision rapid improvements to come. Students are rapidly developing complex prompt engineering skills to accomplish their tasks. Students are being pushed to use generative AI in their co-curricular and professional activities and they are drifting into using it in their academic activities, thinking it through less than one might think, given their concerns. Students' sophistication in appraising generative AI may be part of the problem: they are certain that they can benefit from AI's potential while easily avoiding AI's pitfalls. Many students who end up in my office should not have been so sure.

Messeri and Crockett break down the juncture of communities of knowledge and the individual use of AI. Again, this is an issue that long predates AI, which is exacerbated by it. We have at our disposal vast pockets of knowledge of which we lack proficiency (if you doubt this, ask yourself how a toilet works). Researchers learned as students how to access trustworthy sources of the knowledge they require but do not possess. The anthropomorphized AI exudes a confident tone, measured descriptions, and well-placed caveats to make it seem like a trusted partner in our understanding; this tone is engineered. The "marketing hype" portrays AI "partners" as more accomplished and more trustworthy than human experts. They are engineered to produce "broad, reductive, and quantitative results" that are reassuring in addition to being useful. The assuring quality is crucial for a return on the vast investment in these tools and the source of the greatest epistemic risk. The risk is a confluence of elements: AI tools are consistently made to seem more credible than they actually are; when people perceive that there is a widespread understanding of a particular topic they have shown to have an overestimation of their own knowledge. The direct correlation between the two creates an illusion of understanding, with researchers believing that both they and the AI understand more than they actually do.

The risk to individual studies is obvious, but hundreds of researchers consuming and circulating each other's scholarship could lead to an epistemological collapse.

I want to end by asking what our worry is with the use of generative AI. My guess is that we are worried about a new kind of plagiarism. But that begs the question why we are worried about the old kind of plagiarism. Particularly with the new AI version, there is usually no specific author whose work is being kidnapped. Plagiarism is commonly understood as theft, I don't think that's quite right. Plagiarism is better understood as the failure to engage, to seem as though you are engaging a scholarly conversation when you are not. We best begin those conversations with humility, expecting that our initial position will contain some fundamental misunderstanding, the correction of which leads to discovery and new knowledge. My worry is that the sheen of competence in AI generated answers creates a powerful illusion of understanding. In the thrall of that lovely illusion, how will we remind ourselves that behind every answer there is always something we don't know?