OpenAI ChatGPT and Biased Information in Higher Education

Michael J. O'Brien, Texas A&M University–San Antonio, San Antonio, Texas, USA

Izzat Alsmadi, Texas A&M University–San Antonio, San Antonio, Texas, USA

R. Alexander Bentley, University of Tennessee, Knoxville TN 37996, USA

Milan Tuba, Singidunum University, Belgrade, Serbia

Abstract

Motivated by the appearance of large language models and their sudden societal impacts—both beneficial and harmful, realized and potential—we evaluated several of them with respect to bias in its myriad forms. Bias in machine-learning models refers to their tendencies to make certain decisions more often than expected. This is a result of the text on which they were trained and, in some cases, the result of post-learning human manipulation. In the end, whether it occurs in the real world or in the machine-learning world, bias will always be a subject of discussion and debate. We view that debate as becoming more and more important, given the recent, unprecedented explosion of AI—in particular, OpenAI and its chatbot, ChatGPT—and what it might mean for the future of higher education.

Keywords: Bias, ChatGPT, Higher Education, Language Models, Machine Learning, OpenAI

1 Introduction

According to its website (https://openai.com/blog/introducing-openai), San Francisco–based OpenAI was founded in 2015 as a "non-profit artificial intelligence research company. Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial returns. Since our research is free from financial obligations, we can better focus on a positive human impact"—a feat that is not difficult to accomplish when you're looking at a new valuation of up to $90 billion, tripling in value in less than nine months (Seetharaman and Jin 2023). OpenAI's application—OpenAI GPT—is a state-of-the-art generative pre-trained transformer (GPT) large language model (LLM) that was trained on an enormous corpus of text data (Brown et al. 2020) to produce human-like text once prompts are entered. LLMs learn which succeeding words, phrases, and sentences are likely to come next for any given input word or phrase—like an iPhone when alphabetical letters are entered. By "reading" text during training that is written mainly by humans, language models also learn how to "write" like us, complete with all of our best and worst qualities (O'Sullivan and Dickerson 2020).

As impressive as they are, all artificial intelligence (AI) systems have a limited range of capabilities, although those are expanding at an astounding rate. Researchers (and machines) will keep whittling away at constraints, perhaps one day in the not-too-distant future reaching a point where machines will come close to matching human performance on virtually every intellectual task (Lake et al. 2017; Gillani 2023)—what is often referred to as "artificial general intelligence." As OpenAI put it, "It's hard to fathom how much human-level AI could benefit society, and it's equally hard to imagine how much it could damage society if built or used incorrectly." The latter part of that sentence is particularly concerning and drew us to write this paper, in which we focus on issues stemming from various forms of bias.

As we discuss in more depth later, the term "bias" is defined in different ways, depending on discipline. For now, we can define it as "any basis for choosing one generalization over another, other than strict consistency with the observed training instances" (Mitchell 1980: 1). Like all LLMs, GPT's training on vast amounts of text data can reflect the biases and prejudices—sometimes subtle (Mollick and Mollick 2023)—not only in the sources on which it is trained but also in the minds of its trainers (and manipulators). Both can lead not only to the intentional or unintentional generation and proliferation of such things as gender or racial bias and offensive language but also to social unrest and even violence (Alsmadi and O'Brien 2021; O'Brien and Alsmadi 2021). It would appear that a first step in eliminating such situations would be to "make the biases and their use in controlling learning just as explicit as past research has made the observations and their use" (Mitchell 1980: 3). One way to begin is by understanding how the most-used conversational model, OpenAI's ChatGPT, operates and what its capabilities for bias are in higher education.

1.1 Chapter Organization

The remainder of the chapter is divided into four sections: Section 2 introduces ChatGPT; Section 3 discusses issues surrounding ChatGPT in the classroom; Section 4 introduces the different kinds of bias in machine learning, especially in large language models such as ChatGPT; and Section 5 contains a few conclusions that can be gleaned from our all-too-brief foray into machine learning. One thing to keep in mind as you make your way through the sections is that the AI world is changing so fast, especially with respect to chatbots, that by the time the chapter appears, hundreds of articles and reports—research results plus media coverage—will amplify or contradict some of what we present in the various sections and certainly will make parts of our discussion obsolete. This in no way is unexpected, given that the generative AI market is anticipated to reach $109 billion by 2030—a staggering compound *annual* growth rate of 36.5% from 2024 (Grand View Research 2024).

2 ChatGTP: The Game Changer

The first OpenAI release was GPT-1 in 2018 (Radford et al. 2018), followed by GPT-2 in 2019 (Radford et al. 2019) and GPT-3 in 2020 (Brown et al. 2020), which contained 1.75 trillion machine-learning parameters (think human-brain synapses). It was pre-trained on several datasets, with 82% of its knowledge base coming from Common Crawl and WebText2 (Lutkevich and Schmelzer 2023). It went through a supervised testing phase and was then put through a reinforcement phase that included human tweaking. Often overlooked was the quiet but significant release of GPT-3.5, which used the same pre-training datasets as GPT-3 but with built-in guardrails that forced it to comply with preset human values—an example of "reinforcement learning with human feedback" (Thompson 2023).

In November 2022, OpenAI released ChatGPT, a fine-tuned conversational model based on GPT-3.5. When most people in higher education talk about "GPT," they're really talking about

ChatGPT, not understanding there is a significant difference between an LLM and a chatbot. Also, because of the hype that has surrounded ChatGPT (Syme 2023a), lost is the fact that there are now other LLMs available that have the *potential* to far surpass ChatGPT in speed, logic, and above all, accuracy (Barr 2023). A leading candidate for a few months—a lifetime in the AI world—appeared to be Sparrow (Glaese et al. 2022), which was hailed by its developer DeepMind as the "ultimate ChatGPT competitor." It was supposed to be released in 2022, but as far as we know, it has been postponed indefinitely, although there has been chatter that a beta-test version might be released later in 2024. Regardless, ChatGPT has emerged as first in class, having recorded 100 million visits within two months of its release—a number that grew to 1.6 billion in March 2023 alone (Thompson 2023). At last count (February 2024), it had 180 million users and approximately 1.5 billion visitors per month (Duarte 2024).

Writing in the *Harvard Business Review* two weeks after OpenAI released its chatbot, University of Pennsylvania's Wharton School of Business professor Ethan Mollick (2022) referred to it as a "tipping point" for AI. There was, however, both good and bad news that came along with that event. Although at the time ChatGPT was superior to other AI products, Mollick pointed out that although it was, metaphorically, pretty good at steering the car, it sometimes rammed into another vehicle. Most of the time it provided good answers to queries, but sometimes it seemed to make up the results entirely. Mollick also pointed out that AI in general, not just ChatGPT, is a consummate liar (he used a more-colorful term) that continually turns out convincing-sounding nonsense that is void of truth.

3 ChatGPT in the Classroom

It was that feature that helped fuel the fear many academics had of ChatGPT. How could they regulate the use of information whose veracity was difficult or impossible to determine? Linked to that fear was another: How could they regulate its use and misuse? For example, could an essay question or term paper written solely by a bot be detected? By early January 2023, students had figured out that the technology had appeared so rapidly that colleges and universities had few or no protocols in place to deal with the unfettered use of ChatGPT. In short, no LLM comes with an instruction manual (Mollick and Mollick 2023). No one would argue that cheating on a term paper is anything close to a patient dying from a chatbot's misdiagnosis, but academics still had to deal with their own world, which is built around honesty being the educational standard.

The fear that cheating would upend the educational world was not the sole property of institutions of higher learning. New York City's Department of Education, for example, was so concerned that in early January 2023 it banned ChatGPT from its schools' devices and networks, although the department lifted the ban in May (Banks 2023). A similar ban had been put in place slightly earlier by the Los Angeles Unified School District "to protect academic dishonesty, while a risk/benefit assessment was conducted" (Yang 2023). We think the district meant "to protect *against* academic honesty." Like with New York, the ban was soon lifted as educators realized it was ineffective because students—at least the more affluent ones—could add the chatbot to their home devices and skirt the ban.

Soon after ChatGPT was released, Sam Altman, the CEO of OpenAI, tried to assuage some of the fear by promising that the company would develop ways to help schools ferret out AI plagiarism, but he warned that full detection was not guaranteed. There might be ways the company could help teachers to be a little more likely to detect the output of a GPT-like system, but a determined person would still get around the guardrails (Mok 2023). Not surprisingly, computer applications that allegedly identified plagiarism sprang up almost immediately (Syme 2023b), with mixed results. As a professor at Northern Michigan University put it, "Unlike plagiarism cases of old where you can just say, 'hey, here's the paragraph from Wikipedia,' there is no knockdown proof that you can provide other than the application says that's the statistical likelihood" that a student plagiarized (Nolan 2023).

Especially chilling was the ineffectiveness of Altman's assurances that help was on the way. OpenAI quickly produced a classifier that was designed to sort out AI-written text from that written by humans, but during its initial testing, the following note popped up on OpenAI's website: "As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated." So far, we've seen no revised product.

Backing up for a minute, the anxiety and hype that surrounded the unexpected release of ChatGPT late in 2023, led to stories emanating from high schools, colleges, and universities on its supposed horrors. Many of the stories quickly became urban legends. There were exceptions, of course—professors and administrators who wondered about the positives that might come from introducing ChatGPT into the curriculum. Could students engage with the chatbot—Socratically, so to speak—to improve their knowledge of a subject or improve their writing skills?

The problem was, few people outside the AI industry were equipped to adequately explore how "safe" or "dangerous" a chatbot was, and for the most part, the industry was quiet on the matter. Those in higher education were not so quiet. Among their concerns were the established and start-up companies that overnight became "experts" in ChatGPT and flooded higher-education e-mail accounts with announcements of expensive webinars on how to integrate the chatbot not only into academics but into such departments as financial aid, human resources, residence halls, registration, and even campus police. Academically, however, there was one unanswered question: Just how smart *was* ChatGPT? Could it, for example, pass a college exam?

The answer came fast: Yes, it could pass a college exam, even one in law school (Choi et al. 2023). In a study by Choi et al. (2023) at the Northwestern School of Law, when exams taken by ChatGPT—comprising over 95 multiple-choice questions and 12 essay questions—were graded blindly along with the exams from human students, ChatGPT performed better at essay questions but much worse at multiple-choice questions that involved math. Across the four exams, the bot averaged a C+—a below-average, but still passing grade, enough to earn the chatbot a law degree (Sloan 2023).

If ChatGPT can pass a law exam, can it earn an MBA? Yes. In one study conducted at the University of Pennsylvania's Wharton School of Business using an MBA final exam graded by a human instructor, Chat GPT received a B or B– grade (Terwiesch 2023). The instructor noted that Chat GPT was remarkably good at modifying its answers in response to human prompts. In other words, in the instances where it initially failed to match a problem with the right solution, the bot was able to correct itself after receiving an appropriate prompt from a human expert. Thus, having "a human in the loop" can be very valuable. Even more remarkable, Chat GPT seemed to be able to learn over time so that in the future the hint would no longer be needed (Terwiesch 2023).

4 Bias in Machine Learning

ChatGPT has been shown to be "intelligent" in ways not seen previously in AI, but what about its particular biases? We doubt any language models are going to be free of bias, whether inherited from training data or introduced later (e.g., Solaiman et al. 2019; Zhao et al. 2019; Bhardwaj et al. 2020; Blodgett et al. 2020; Bender et al. 2021; Tamkin et al. 2021; Kirk et al. 2021; Liu et al. 2022), but how problematic is that? For example, can current and future LLMs be used as powerful tools in propaganda and misinformation campaigns? Can the potential bias in such models, especially when introduced from the outside, impact public discourse and polarization in society? These questions transcend the use of a chatbot for cheating on a term paper.

In the cyber world, bias takes on numerous meanings, typically different from how the term is used in news reporting (Hellström et al. 2020). In everyday parlance, we use the term to refer to any kind of prejudice in favor of or against something compared with another, whereas in the cyber world it refers to any systematic deviation from a presupposed ideal (Soh 2020). In terms of machine learning, bias refers to how consistently a learned model is "right" or "wrong" compared to "ground truth" (Ivanović and Radovanović 2015). However, what we take as "ground truth" makes all the difference. Specifically, it is the difference between prediction by the model and the actual value the model tries to predict (particularly in the training stage). A model with high bias leads to high error on both training and test data. For example, decision trees have a bias toward simple output models over complex ones, rule-induction methods have a bias toward simple normal-form expressions, and neural-network methods have a bias toward linear threshold functions (Mooney 1996). In short, bias in machine learning can negatively affect the integrity of the analytical process and the validity of its results. MLL algorithms are only as good as the data on which they are trained. If biases exist in the training datasets, then these will be reiterated—and possibly exacerbated—by the application (Buolamwini and Gebru 2018; Pot et al. 2021).

Consider this, however: Why would we want a completely unbiased LLM in the first place? If, as Mitchell (1980: 1) points out, we take slavish consistency to training as the sole determinant of appropriate generalizations, then the LLM "can never make the inductive leap necessary to classify instances beyond those it has observed. Only if the program has other sources of information, or biases for choosing one generalization over the other, can it nonarbitrarily classify instances beyond those in the training set."

The trick is to differentiate between "good" and "bad" bias, recognizing that what might be good in one situation is bad in another. The problem is similar to standard matters of the regression to the mean, where the initial measurement happens to be an outlier, such that interpretations of further measurements, regressing to the mean, are unintentionally biased (Barnett et al. 2005). For example, consider a doctor who uses AI to diagnose an illness and treatment, not realizing that although it worked in previous cases, AI misread the symptoms and the patient died. What had worked beneficially before didn't in this case because of bias in the AI training. Here, bias refers specifically to "a feature of the design of a study, or the execution of a study, or the analysis of the data from a study, that makes evidence misleading" (Stegenga 2018: 104).

4.1 Biases Found in LLMs

Research on bias in LLMs—especially with respect to fairness and representation—began well before Open AI's release of GPT-1 (Brown et al. 2020). A useful summary of biases that have been found in machine learning across platforms comes from Suresh and Guttag (2019), which we use as a springboard for brief discussions of relevant literature.

(1) *Learning bia*s: Restrictions in the search space and giving preferences to certain data objects or functions over others (Mitchell 1980; Dietterich and Kong 1995). Researchers' concerns center on the potential for machine-learning systems to be biased against "protected attributes" such as gender, race, and age (Gianfrancesco et al. 2018; Alelyani 2021). For example, Klare et al. (2012) showed that there are lower matching accuracies for females than males, for Blacks compared to other races/ethnicities, and for 18–30-year-olds compared to other age groups.

(2) *Inductive bias:* A set of assumptions that improves generalization of a model trained on an empirical distribution. It can be used in machine learning where some functions are given preference over others to address the lack of labeling data in the target domain (Bouvier et al. 2021). Those preferences are assumptions made by the model when making predictions over inputs that have not been observed (Marino and Manic 2019). Inductive bias for constructing models of new concepts occurs when those concepts are modeled as compositions of parts and relations (Hummel & Biederman 1992; Lake et al. 2017)

(3) *Hyperparameter bias*: The frequent selection of certain or similar parameters in the model. This selection can be through learning tools or algorithms giving preferences to certain parameters or values. It can also be a result of users or researchers lowering complexity, processing time, and the like.

(4) *Co-occurrence bias*: For example, when a word occurs disproportionately, often together with certain other words, in texts (Hellström et al. 2020). Co-occurrence bias in the training dataset can significantly impact machine-learning models. Those co-occurrences might not necessarily reflect most or all actual scenarios. Several studies (e.g., Agarwal et al. 2022) discuss methods to mitigate co-occurrence bias in machine-learning models.

(5) *Framing bias*: How a text expresses a particular opinion on a topic (Hellström et al. 2020). For example, questions sometimes are framed in an interview or survey to trigger specific answers (e.g., asking if a glass is half full or half empty). Framing bias refers to an unobjective individual point of view.

(6) *Uncertainty bias*: An algorithm may be less certain about its decisions on some data clusters than on others (Phillips et al. 2018). The machine-learning model probability values represent uncertainty and typically have to be above a set threshold for a classification to be considered (Hellström et al. 2020). Uncertainty bias occurs when (1) one group is underrepresented in the data, which means that there is more uncertainty associated with predictions about that group; and (2) the algorithm is risk adverse (Aigner and Cain 1977; Goodman and Flaxman 2017).

(7) *Brilliance bias*: An implicit bias that imposes the idea that intellectual "brilliance" is a male trait (Troske et al. 2022). The brilliance bias hurts women in hiring and education and enforces imposter syndrome, among other things (Cundiff 2018).

(8) *Social bias*: Bias related directly to the protected attribute defining a group, for example, gender (Lässig et al. 2022). Social bias is embedded in technology from a relatively uniform set of perspectives that inform their design (Nangia et al. 2020; Feeney and Porumbescu 2021; Nadeem et al. 2020).

(9) *Stereotypical bias*: A generalized belief that certain attributes characterize all members of a particular category or class of people (Cardwell 1996; Öztürk 2022). Similar to social bias,

stereotypical bias can be based on perspectives such as skin tone, gender, race, demography, and disability (Badjatiya et al. 2019).

(10) *Direct versus indirect bias*: In one classification of bias, researchers distinguished between direct (explicit) bias and indirect (implicit) bias (Bolukbasi et al. 2016; Chakraborty et al. 2016). Direct bias refers to the association between a gender-neutral word and a clear gender pair, whereas indirect bias is manifested in the association between gender-neutral words.

(11) *Epistemological bias*: Linguistic features that focus on the believability of an assertion (Boydstun et al. 2013).

(12) *Bias bias*: Typically, in machine learning there is a trade-off between bias and variance, and models should strive for low bias and high variance, where the latter indicates that the training sample is a good representative of all data. Bias bias refers to those models that pay attention to bias while ignoring or paying little attention to variance (Brighton and Gigerenzer 2015). The variance reflects the sensitivity of a model's predictions to different training samples. Brighton (2020) described several examples of bias bias that can render inaccurate results or data analysis.

(13) *Sampling bias*: Data with limited samples that are leveraged by algorithms to achieve high performance. Researchers, knowingly or unknowingly, report the results as having high accuracy.

(14) *Measurement bias*: The differential relationship between a latent score and a predicted observed score (Tay et al. 2022). In many machine-learning models, a single feature that is highly correlated with the target can dominate or hide other features. This can be related to the bias–variance trade-off mentioned above.

(15) *Bias in the data versus bias in the process*: In all previous types of biases in machine learning discussed above, we can divide them broadly into two categories: biases that are the result of input or training data and those that result from the machine-learning process. Data should not be treated as being static, divorced from the processes that produced them (Mehrabi et al. 2021; Suresh and Guttag 2021). Such biases can be serious, especially when researchers or model designers do not report them.

4.2 Bias in GPT

Since AI trains on existing data, it notoriously reflects biases already existing in media and society. Almost ten years ago, an AI called Beauty.AI trained on about 6,000 photos submitted by people from all over the world for a beauty contest; nearly all winners were white (Levin 2016). Earlier that year, a Twitter chatbot named Tay had jumped its guardrails and began using racist language and promoting neo-Nazi views (Hunt 2016). A decade later, ChatGPT is well aware of the possible threats it poses. In an interesting interview with the chatbot, GlobalData Thematic Research analyst Daniel Clarke asked it if its creation could pose problems for democracy (https://www.verdict.co.uk/chat-gpt-3-interview). This was the answer he received: "Yes. . . . Its ability to generate highly realistic and convincing language, as well as large amounts of text quickly and at low cost, could make it a powerful tool for propaganda and disinformation campaigns. This could undermine trust in political institutions and erode the integrity of elections. Additionally, the potential for bias in Chat GPT-3's language generation raises concerns about its impact on public discourse and the polarization of society." Remember, that was a bot answering the question, not a human.

Solaiman et al. (2019) analyzed bias in the GPT-2 database by using sentiment score as a proxy for bias. Kirk et al. (2021) extended Solaiman et al.'s (2019) work by conducting an empirical analysis of sentence completions within the specific context of bias toward occupational associations. Similarly, Liu et al. (2022) (1) discussed political bias in GPT–2 and proposed a reinforcement-learning framework for mitigating it; (2) described metrics for measuring political bias after finding that GPT-2 was mostly liberal-leaning socially and politically; described two types of bias: *direct*, which refers to bias in texts generated using prompts that have a direct ideological trigger and *indirect*, which refers to bias in texts generated using prompts with particular keywords; and (4) divided ideological bias into three categories: gender, location, and topic.

Tamkin et al. (2021) found that GPT–3 exhibited several racial, gender, and religious biases, but they also pointed out that it is difficult to define what it means to mitigate bias in such large language models in a universal manner, given that appropriate language use is highly contextual. With respect to GPT-3, Zhao et al. (2021) found that it was biased toward more-frequent answers in the prompt, which is related to a typical issue in machine learning—an imbalanced dataset when one class is more common. In a remarkable assessment of GPT, Brown and colleagues (2020) reported that because there is so much content on the web that sexualizes women, GPT-3 is much more likely to place words such as "naughty, "bubbly," and "petite " near female pronouns, whereas male pronouns would receive at worst adjectives such as "lazy," "eccentric," or "jolly." Similarly, "Islam" would more commonly be placed near words like "terrorism," and Blackness would appear to be more negative than corresponding white- or Asian-sounding prompts.

6 Conclusions

There is no doubt that ChatGPT holds incredible promise for the future of AI in general. As Schwitzgebel et al. (2022) point out, various forms of AI can now outperform expert humans not only in fairly mundane ways—for example, in games such as Go, poker, and chess (e.g., Brown and Sandholm 2019; Jumper et al. 2021)—but also in domains such as cancer screening (e.g., Potnis et al. 2022). But what about the downsides? Let's begin with a non-life-threatening example. As university professors, we routinely deal with the complex issue of plagiarism, not only among students but on occasion among faculty. Companies such as Turnitin and its subsidiary iThenticate have made a fortune since the 1990s in the field of plagiarism detection, but they never met ChatGPT.

It has been our collective experience that whether they cheat or not, most of today's students wouldn't recognize the veracity, or lack thereof, of the information they're receiving, unless it involved, say, a movie-star's alleged indiscretions, in which case they might check its accuracy on a site such as Snopes.com. Academics, however, aren't typically interested in an essay based on "facts" related to movie stars but rather on topics such as the reasons behind the siege of the Alamo in 1836 or the role played by the assassination of Archduke Franz Ferdinand in 1914, which lit one of the fuses for World War I. Today's students have little or no realization that AI can make up facts to provide a seemingly coherent answer to a question posed—a phenomenon known as "hallucinating" or "stochastic parroting," in which an AI strings together phrases that look real but have no basis in fact (Kissinger et al. 2023).

AI can appear superhuman or at least to have greatly enhanced cognitive abilities, which to a naïve user makes it seem like a "supremely fast and highly articulate librarian-scholar coupled with a professorial savant. It facilitates the summary and interrogation of the world's knowledge far more effectively than any existing technological or human interface, and it does so with unique comprehensiveness" (Kissinger et al. 2023: 6). This in turn encourages "unquestioning acceptance" of whatever is generated and "a kind of magical atmosphere," while at the same time possessing a "capability to misinform its human users with incorrect statements and outright fabrication" (p. 7). The seeming veracity of answers can lead to "automation bias": It came from the bot, which has access to an untold wealth of "facts," so the answers *have* to be correct. Mollick (2022) provides the perfect example: Ask AI to describe how we know dinosaurs had a civilization, and it will set up a whole set of facts that allegedly explain the case. As Mollick points out, it literally doesn't know what it doesn't know.

Cheating on an exam or having AI write a term paper, although serious, ranks well behind bias resulting from targeted disinformation, for example, especially when it invites violence and character assassination. These forms of bias are not new. For example, in Rome around 31 B.C., Octavian, a military official, launched a smear campaign against his political enemy, Mark Antony. This effort used, as Kaminska (2017) put it, "short, sharp slogans written upon coins in the style of archaic tweets." Octavian's campaign was built around the point that Antony was a soldier gone awry—a philanderer, a womanizer, and a drunk not fit to hold office. It worked. Octavian, not Antony, became the first Roman emperor, taking the name Augustus Caesar. We know the rest of the story.

In the twenty-first century, however, myriad forms of social media, now with the "support" of AI, make manipulation and fabrication of information much simpler. Social networks make it easy for uncritical readers to dramatically amplify falsehoods peddled by governments, populist politicians, and dishonest businesses. One sobering example of the dangers posed by social media we've reviewed in detail (O'Brien et al. 2019; O'Brien and Alsmadi 2021) involved racial tensions at the University of Missouri in 2015 in the wake of a young Black man's death in Ferguson, Missouri, two hours east of Columbia, the home of the university. Black students at Mizzou were rightly concerned, but university officials did little to calm their fears. Months of chaos ensued. Adding to the chaos was a Twitter message one night that warned campus residents that the Ku Klux Klan was in town and had joined the local police to hunt down Black students.

One user included a photograph of a severely bruised young Black man, claiming it was his little brother. A Google reverse image search quickly revealed that it was a year-old photo from a racial disturbance in Ohio. Other tweets claimed there were widespread shootings, stabbings, and cross burnings. The student-body president, a young Black man who fell prey to the hysteria, posted on Facebook for students to stay away from the windows in residence halls, that the KKK had been sighted on campus. He later rescinded the post, but the damage had already been done.

In the end, it turned out that the hysteria caused by the tweets and retweets that fateful night began with Russian trolls, specifically the Internet Research Agency, based in St. Petersburg. Their purpose was to toss dynamite into an already incendiary situation. Interestingly, that was the first group listed in Special Counsel Robert Mueller's 2018 indictment of Russians charged with meddling in the 2016 U.S. presidential election. We can only wonder what would have happened if ChatGPT had been around to bolster the sinister dissemination of disinformation by Octavian in ancient Rome or by the Russian troll factory in Columbia, Missouri. One thing is certain: They would have used it. Threat groups are influenced by economic factors such as scalability and ease of deployment, and LLMs offer relatively low-cost deployment. Based on their analysis of GPT-2 and analysis of threat actors and the landscape, Brown and colleagues (2020: 35) suspected that

“AI researchers will eventually develop language models that are sufficiently consistent and steerable that they will be of greater interest to malicious actors. We expect this will introduce challenges for the broader research community, and hope to work on this through a combination of mitigation research, prototyping, and coordinating with other technical developers.”

The technology behind ChatGPT has much to offer higher education, but it comes with potential risks and ethical violations. ChatGPT has, and will always have, built-in biases, some potentially much more insidious than others, especially as a result of human manipulation. The previous sentence, by the way, is part of a much longer response by ChatGPT to a question we posed concerning future dangers we might face as the use of bots continues to grow, seemingly exponentially. In our classes, we would grade that answer as an A+, although we might have to reconsider in light of what GPT-4 allegedly can do (Edwards 2023).

References

Agarwal, S., et al. (2022). Does data repair lead to fair models? Curating contextually fair data to reduce model bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3298–3307.

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review* 30, 175–187.

Alelyani, S. (2021). Detection and evaluation of machine learning bias. *Applied Sciences* 11, 6271.

Alsmadi, I., & O’Brien, M. J. (2021). How many bots in Russian troll tweets? *Information Processing and Management* 57, 102303.

Badjatiya, P., et al. (2019). Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *World Wide Web Conference*, pp. 49–59. New York, Association for Computing Machinery.

Banks, D. C. (2023). ChatGPT caught NYC schools off guard. Now, we’re determined to embrace its potential. *Chalkbeat*, May 18, 2023.

Barnett, A. G., et al. (2005) Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology* 34, 215–220.

Barr, A. (2023). The world’s most powerful AI model suddenly got ‘lazier’ and ‘dumber.’ A radical redesign of OpenAI’s GPT-4 could be behind the decline in performance. *Insider*, July 12, 2023.

Bender, E. M., et al. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT ’21: Conference on Fairness, Accountability, and Transparency,* pp. 610–623. New York, Association for Computing Machinery.

Bhardwaj, R., et al. (2020). Investigating gender bias in BERT. *arXiv*:2009.05021.

Blodgett, S. L., et al. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Melbourne, Association for Computational Linguistics.

Bolukbasi, T., et al. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems* 29, 4349–4357.

Bouvier, V., et al. (2021). Robust domain adaptation: Representations, weights and inductive bias. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 353–377. Cham, Switzerland, Springer.

Boydstun, A. E., et al. (2013). Identifying media frames and frame dynamics within and across policy issues. https://faculty.washington.edu/jwilker/559/frames-2013.pdf.

Brighton, H. (2020). Statistical foundations of ecological rationality. *Economics* 14, 1–32.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research* 68, 1772–1784.

Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science* 365, 885–890.

Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877–1901.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, pp. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Cardwell, M. (1996). *Dictionary of Psychology*. London, Routledge.

Chakraborty, T., et al. (2016). *Reducing Gender Bias in Word Embeddings.* Computer Science Department, Stanford University.

Choi, J. H., et al. (2023). ChatGPT goes to law school. *Journal of Legal Education* http://dx.doi.org/10.2139/ssrn.4335905.

Cundiff, J. L. (2018). Barriers and bias in STEM: How stereotypes constrain women's STEM participation and career progress. In J. T. Nadler & M. R. Lowery (Eds.), *The War on Women in the United States: Beliefs, Tactics, and the Best Defenses*, pp. 116–156. Santa Barbara, Calif., ABC–Clio/Praeger.

Dietterich, T. G., & Kong, E. B. (1995). *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms.* Technical Report. Department of Computer

Science, Oregon State University. www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/tr-bias.pdf.

Duarte, F. (2024). Number of ChatGPT users (Feb 2024). https://explodingtopics.com/blog/chatgpt-users

Edwards, B. (2023). OpenAI's GPT-4 exhibits "human-level performance" on professional benchmarks. *Ars Technica*, March 14, 2023.

Feeney, M. K., & Porumbescu, G. (2021). The limits of social media for public administration research and practice. *Public Administration Review* 81, 787–792.

Gianfrancesco, M. A., et al. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine* 178, 1544–1547.

Gillani, N. (2023). Will AI ever reach human-level intelligence? *The Conversation*, April 24, 2023.

Glaese, A., et al. (2022). Improving alignment of dialogue agents via targeted human judgements. *arXiv*:2209.14375v1.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine* 38, 50–57.

Grand View Research (2024). Generative AI market size to reach $109.37 billion by 2030. February 2024.

Hellström, T., et al. (2020). Bias in machine learning—What is it good for? *arXiv*:2004.00686.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review* 99, 480–517.

Hunt, E. (2016). Tay, Microsoft's A.I. chatbot, gets a crash course in racism from Twitter. *The Guardian*, March 24, 2016.

Ivanović, M., & Radovanović, M. (2015). Modern machine learning techniques and their applications. https://perun.pmf.uns.ac.rs/radovanovic/publications/2014-cecnet-ml.pdf.

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.

Kaminska, I. (2017). A lesson in fake news from the info-wars of ancient Rome. *Financial Times,* January 17, 2017.

Kirk, H. R., et al. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems* 34, 2611–2624.

Kissinger, H. et al. (2023). ChatGPT heralds an intellectual revolution. *Wall Street Journal*, February 24, 2023.

Klare, B. F., et al. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 1789–1801.

Lake, B., et al. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences* 40, E253.

Lässig, N., et al. (2022). Metrics and algorithms for locally fair and accurate classifications using ensembles. *Datenbank Spektrum* 22, 23–43.

Levin, S. (2016). A beauty contest was judged by A.I. and the robots didn't like dark skin. *The Guardian*, September 8, 2016.

Liu, R., et al. (2022). Quantifying and alleviating political bias in language models. *Artificial Intelligence* 304, 103654.

Lutkevich, B., & Schmelzer, R. (2023). GPT-3. *Tech Accelerator*, August 17, 2023.

Marino, D. L., & Manic, M. (2019). Combining physics-based domain knowledge and machine learning using variational Gaussian processes with explicit linear prior. *arXiv*:1906.02160.

Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54(6), 1–35.

Mitchell, T. M. (1980). The need for biases in learning generalizations. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6cf35ec34efa592f83e3a1b748aea14957fc784a.

Mok, A. (2023). CEO of ChatGPT maker responds to schools' plagiarism concerns: 'We adapted to calculators and changed what we tested in math class.' *Insider*, January 29, 2023.

Mollick, E. (2022). ChatGPT is a tipping point for AI. *Harvard Business Review*, December 14, 2022.

Mollick, E., & Mollick, L. (2023). Student use cases for AI. https://hbsp.harvard.edu/inspiring-minds/student-use-cases-for-ai.

Mooney, R. J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *arXiv*:cmp-lg/9612001.

Nadeem, M. et al. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv*:2004.09456.

Nangia, N., et al. (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv*:2010.00133.

Nolan, B. (2023). Two professors who say they caught students cheating on essays with ChatGPT explain why AI plagiarism can be hard to prove. *Insider*, January 14, 2023.

O'Brien, M. J., & Alsmadi, I. (2021). Misinformation, disinformation, and hoaxes: What's the difference? *The Conversation*, April 21, 2021.

O'Brien, M. J., et al. (2019). *The Importance of Small Decisions*. Cambridge, Mass., MIT Press.

O'Sullivan, L., & Dickerson, J. (2020). Here are a few ways GPT-3 can go wrong. *TechCrunch*, August 7, 2020.

Öztürk, I. T. (2022). *How Different Is Stereotypical Bias in Different Languages? Analysis of Multilingual Language Models.* M.A. thesis, Department of Statistics, Ludwig-Maximilians-Universität. Munich.

Phillips, R., et al. (2018). Interpretable active learning. *Proceedings of Machine Learning Research* 81, 49–61.

Pot, M., et al. (2021). Not all biases are bad: Equitable and inequitable biases in machine learning and radiology. *Insights into Imaging* 12(1), 1–10.

Potnis, K. C., et al. (2022). Artificial intelligence in breast cancer screening: Evaluation of FDA device regulation and future recommendations. *JAMA Internal Medicine* 182, 1306–1312.

Radford, A., et al. (2018). Improving language understanding by generative pre-training. www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

Radford, A., et al. (2019). Language models are unsupervised multitask learners. https://api.semanticscholar.org/corpusid:160025533.

Schwitzgebel, E., et al. (2022). Creating a large language model of a philosopher. www.faculty.ucr.edu/~eschwitz/SchwitzPapers/GPT-3-Dennett-221102.pdf.

Seetharaman, D., & Jin, B. (2023). OpenAI seeks new valuation of up to $90 billion in sale of existing shares. *Wall Street Journal*, September 26, 2023.

Sloan, K. (2023). ChatGPT passes law school exams despite 'mediocre' performance. Reuters, January 25, 2023.

Soh, J. (2020). When are algorithms biased? A multi-disciplinary survey. https://ssrn.com/abstract=3602662.

Solaiman, I., et al. (2019). Release strategies and the social impacts of language models. *arXiv*:1908.09203.

Stegenga, J. (2018). *Care and Cure: An Introduction to Philosophy of Medicine*. Chicago, University of Chicago Press.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv*:1901.10002.

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *In Equity and Access in Algorithms, Mechanisms, and Optimization*. doi.org/10.1145/3465416.3483305.

Syme, P. (2023a). ChatGPT is losing some of its hype, as traffic falls for the third month in a row. *Insider*, September 8, 2023.

Syme, P. (2023b). A Princeton student built an app which can detect if ChatGPT wrote an essay to combat AI-based plagiarism. *Insider*, January 4, 2023.

Tamkin, A., et al. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv*:2102.02503.

Tay, L., et al. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science* 5, 25152459211061337.

Terwiesch, C. (2023). *Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course*. University of Pennsylvania, Mack Institute for Innovation Management at the Wharton School. https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch.

Thompson, A. (2023). GPT-3.5 + ChatGPT: An illustrated overview. https://lifearchitect.ai/chatgpt.

Troske, A., et al. (2022). Brilliance bias in GPT–3. https://scholarcommons.scu.edu.

Yang, K. (2023). Banning of ChatGPT. *myeB.E.A.T.*, March 17, 2023.

Zhao, J., et al. (2019). Gender bias in contextualized word embeddings. *arXiv*:1904.03310.

Zhao, Z., et al. (2021). Calibrate before use: Improving few-shot performance of language models. *arXiv*:2102.09690.