TEACHING CRITICAL AI LITERACY:
Advice for the New Semester

Prepared by Lauren M. E. Goodlad and Sharon Stoerger in collaboration with the AI Round Table Advisory Council and with support from the Office of Teaching Evaluation and Assessment Research.

This webpage offers a provisional set of resources to help instructors make informed decisions and equip themselves for productive discussions as they prepare for a new semester. We provide

(1) a brief introduction to "artificial intelligence," followed by
(2) discussion of critical AI literacy for educators and students,
(3) discussion of the implications of "generative AI" tools for academic integrity,
(4) suggestions for updating syllabi, and
(5) a short list of potential resources.

"AI" is a complicated subject with many far-reaching implications: we have sought to strike a balance between brevity and comprehensiveness.

1. *What is Artificial Intelligence* (AI)?

Artificial Intelligence (AI) has become a common term for an emerging set of computer technologies that affect individuals, communities, societies, and the environment at a global scale. Although the phrase "AI" was coined in the 1950s, it has undergone multiple transformations as a field of research and, until recently, was familiar to the general public largely as a theme for science fiction.

AI research returned to public discussion in the 2010s when a number of innovations in "deep learning" became possible—largely because of the availability of human-generated data on the internet and through networked devices at an unprecedented scale. At around the same time, these technologies began to power widespread applications including voice assistants, recommendation systems, and automated driver assistance. When technologists speak of "deep learning" (DL), which is a type of "machine learning" (ML), the *learning* in question denotes a computer model's ability to "optimize" for useful predictions while "training" on data through updating the weights in an elaborate set of statistical calculations. The learning is *deep* because of the multiple computational layers in the very large models that DL involves.

The most heavily promoted forms of "AI" today are large language models (LLMs) such as Open AI's ChatGPT, Google's Bard, and Anthropic AI's Claude 2.[1] All of these systems are multi-layered ("deep") statistical models that predict probable word sequences in response to a prompt even though they do not "understand" language in any human-like sense. Through the intensive mining, modeling, and memorization of vast stores of data "scraped" from the internet, reinforced by vast bodies of human gig workers, text generators synthesize a few paragraphs at a time which resemble writing authored by humans. This machine-generated text is not directly "plagiarized" from some original, and it is usually grammatically and syntactically well-crafted.

However, machine-generated content is often factually incorrect. Moreover, when users prompt an LLM to provide sources for information, the cited sources may be wrong or completely fabricated. Likewise, while chatbots are sometimes marketed as question-answering systems (like Apple's Siri) and as effective replacements for search engines, LLMs are *pre-trained* models that do not search the web.[2] The novelist and technologist Ted Chiang has likened a chatbot's compressed model of its training data to a "blurry JPEG." A related limitation to consider is that LLMs are expensive to retrain: for example, although OpenAI's GPT-4 was released in March 2023, its original training involved textual data gathered through 2021. This means that an LLM's training data may not include the most recent developments or insights in various fields.

Since the training data for LLMs is enormous–constituting most of the "scrapable" internet–as models have grown successively larger, their predictive capacities have extended beyond the generation of human-like text: for example, they can now answer some questions in basic math (though still make errors on simple tasks such as three-digit multiplication); in the hands of knowledgeable programmers, they can aid in the writing of computer code. Large image models (LIMs) which are trained on visual images scraped from the internet, are text-to-image generators that respond to a user's textual prompt with a visual output predicted to correspond accordingly. Because these commercial models are proprietary, researchers have no means of determining precisely what training data was involved, what tasks were prioritized for human reinforcement (and under what working conditions), or how much energy and water is required for training and use.

---

[1] Though sometimes described as a "start-up," OpenAI is valued at about $30 billion and has been funded partly through multi-billion dollar investments from Microsoft in exchange for a 49% stake. Anthropic AI was founded by OpenAI employees who disagreed with the direction being taken by OpenAI. Meta's most updated current language model, LLaMA 2, is open source. BLOOM, another open source LLM, was created as a collaboration between more than 1000 researchers.

[2] See, for example, the *MIT Review*'s criticism of replacing search engines with chatbots and, for a research paper on the topic, Shah and Bender (2022). On the various technical challenges to incorporating chatbots into search engines (as in Microsoft's Bing and Google's Bard) see, for example, Vincent (2023).

These data-driven technologies are now collectively designated as "generative AI" and their impressive affordances are sometimes talked up as if they were both "god-like" and likely to lead to imminent catastrophe.[3] This peculiar tension between the need to regulate new technologies in order to reduce their harms, and the tendency of some prominent voices to emphasize so-called existential risks at the expense of actually existing harms is one of several ways that "hype" about chatbots and other predictive "AI" systems complicates the educating of students and the public at large. At Rutgers, the Critical AI @ Rutgers initiative is among many groups, national and international, which hold that promoting *critical AI literacy* is the best possible answer to a dynamic landscape.

2. *What is Critical AI Literacy?*

Chatbots and other modes of generative AI are controversial because of a number of problems that may be hard (or impossible) to eradicate: these include embedded biases, lack of transparency, built-in surveillance, environmental footprint, and more. Teaching *critical AI literacy* thus includes helping students to learn about the existing and potential harms of these tools, whether instructors use them in their teaching or not. To be sure, acquiring in-depth literacy takes time for both educators and students. In the best possible case, students and instructors will learn from each other as they discuss common concerns and experiences.[4]

Below we list the chief concerns about the actually existing harms of chatbots and other "generative" tools and the systems and practices on which they depend.

- *Amplification of Bias, Malignant" Stereotypes, and "Documentation Debt"*: since LLM performance relies heavily on large datasets, the best-performing models are also riddled with bias and stereotypes from largely undocumented data scraped from the internet. Bender, Gebru and colleagues (2021) describe the risks of this "documentation debt" in the context of proprietary datasets that "overrepresent hegemonic viewpoints."[5]

[3] For a point-by-point critique of the "reckless exaggeration" such writing may entail, see Noah Giansiracusa's response to an opinion essay in the *New York Times*.

[4] See Conrad's "Blueprint for An AI Bill of Rights for Educators and Students," for a useful framework for enabling instructors to teach critical AI literacy (including a basis for discussing such "rights" with your students).

[5] Through probing and audits of LLMs, researchers have discovered "persistent toxic" content (Gehmen et al. 2020 3356) and "severe" bias against Muslims (Abid et al. 2021 298). Looking at multimodal models, Birhane and colleagues (2021) have found misogynistic and pornographic content; Hundt and colleagues (2022 753) warn that robots programmed with CLIP (an OpenAI image-to-text classifier), pick up "malignant stereotypes" including "racist, sexist, and scientifically discredited physiognomic behavior." Since LLMs and LIMs rely on correlation and cannot distinguish truth from falsehood or bias from fairness, efforts to mitigate these problematic outputs without human feedback can at best be partial–just as human feedback is itself limited given the size of the models in question. For more recent evidence of untrustworthy model behaviors which await peer review, see Piltch (2023) and Wang et al. (2023). Bender et al. (2021 615) define *documentation debt* as "putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc. While documentation allows for potential accountability, undocumented training data perpetuates harm

- *Copyright Infringement, Lack of Consent, Surveillance, and Privacy Concerns*: the use of copyrighted content scraped from the web without consent for the training of AI models has opened a host of legal questions. (Notably, the *New York Times*, in August 2023, updated its service to forbid use of its content for training data–a new development in practices regarding protection of intellectual property).[6] At the same time, the accumulation of user data during use of commercial chatbots extends the surveillant practices that began with the monetization of social media and search engines, exacerbating data privacy concerns.[7]

- *Environmental Footprint*: because "generative AI" is computationally intensive, the technology uses significantly more energy and water than a simple internet search.[8]

- *Exploitation of Human Labor*: since LLMs are subject to bias, misinformation, and toxicity, the current technology relies on millions of low-paid crowdworkers whose annotating work improves results.[9]

- *Misinformation through "hallucinations," conspiracy theories/misconceptions, and malicious use.*[10]

---

without recourse. Without documentation, one cannot try to understand training data characteristics in order to mitigate some of the" actual and potential harms. For an important study of bias in facial recognition systems see Buoalmwini and Gebru (2018). Foundational research on the topic of algorithmic bias includes Sweeney (2013) O'Neil (2016), Noble (2018), and Benjamin (2019). Broussard's (2019) introduction to AI discusses its cold war-era inception. Research in the field of Artificial Intelligence in Education (AIEd) indicates that AI has the potential to enable beneficial applications in higher education, including intelligent tutoring systems, personalization, and assessment and evaluation (e.g., Luckin and Holmes, 2016); yet it is important to recognize that many of these potential uses have not included a critical reflection of pedagogical research (e.g., Bartolomé, Castañeda, and Adell, 2018; Zawacki-Richter et al, 2019). On related ethical concerns see also Zeide (2023).

[6] A recent essay in the *Atlantic Monthly* documented that hundreds of thousands of copyrighted works are "secretly" being used to train large and proprietary models. On the *New York Times*'s potential legal action against OpenAI, see Allyn (2023).

[7] Zuboff's influential study (2019) describes the underlying business model of tech companies such as Google and Facebook (now Meta) as *surveillance capitalism*. The enormous importance of data accumulation ("big data") in training AI and other digital processes continues to be studied across the disciplines. See, for example, Gitelman (2013), Sadowski (2019), D'Ignazio and Klein (2020), and Denton et al. (2021).

[8] On the environmental footprint of training large models see Strubell et al. (2019) and Luccioni, Viguier, and Ligozat (2023); on the water usage involved in training and prompting chatbots, see Li et al. (2023) for a more holistic discussion of AI's footprint, see Crawford (2021); on the ecological and environmental costs of cloud computing more generally, see, *e.g.*, Hogan and Vanderau (2019) and Monserrate (2022).

[9] Recent journalism documents how supposedly automated chatbots require massive input from low-paid gig workers with the labor of labeling violent and disturbing content often outsourced to low-paid workers in the global South. On the longstanding practice of using human crowdworkers for machine learning and the improvement of automated systems, see, for instance, Ross, Irani et al. (2010), Gray and Suri (2019), and Crawford (2021, chapter 2).

[10] *Hallucination* is the industry's preferred term for the persistent problem of LLMs' confabulation of plausible but factually incorrect responses. As a result of this problem, LLMs sometimes generate nonsensical information such as a wikipedia page for the health benefits of feeding crushed porcelain to babies (see Birhane and Raji 2022). As Klein rightly notes (2023), applying the anthropomorphizing language of "hallucination" to

- *Political Economy, Concentration of Power, Lack of Transparency and Accountability:* the political economy of "AI"' today was forged through the concentration of computing, economic, and data resources in some of the largest and most lucrative companies in the world. Corporations such as Microsoft (and their OpenAI partner) and Google intensively lobby legislators, sometimes "watering down" regulatory demands for transparency and accountability for these dominant companies and their products. Lina Khan, who is chair of the Federal Trade Commission, has described the risks of "AI" in a context of "race-to-the-bottom business models and monopolistic control."[11]

In an educational setting, chatbots also create particular challenges for academic integrity–a subject to which we now turn.

3. Implications for Academic Integrity

The code of conduct at Rutgers states that students must ensure "that all work submitted in a course, academic research, or other activity is the student's own and created without the aid of impermissible technologies, materials, or collaborations." In evaluating policies on "generative" tools, this puts special emphasis on the identification of permissible technologies and the question of whether a given tool impedes the learning goals of the course (including the submission of suitable work that is "the student's own").

At Rutgers, learning goals vary widely across and within schools, disciplines, majors, and levels of difficulty.

For example,

- A computer science instructor teaching an introductory course may wish to prohibit students from using chatbots for coding in order to ensure that they learn

---

a statistical model is misleading and problematic. A second source of misinformation is the mimicking of human-generated falsehoods, misconceptions, and conspiracy theories gleaned from unvetted training data (see Lin, Hilton, and Evans 2022). Both problems explain why human interventions at massive scale have become necessary for making AI systems more reliable and less toxic. A third difficulty is the ease with which humans bent on malicious use can circumvent weak guardrails to generate harmful content at scale. Although "jailbreaking" ChatGPT has become a comical pastime, malicious use of AI systems (which may include deep fakes, fake pornography, and the hacking of cars) is a serious matter that extends far beyond the subject of LLMs per se. See Maiberg (2023) for a disturbing account of how generative models are used to "produce any kind of pornographic scenario…trained on real images of real people scraped without consent from every corner of the internet."

[11] See Whittaker (2021: 51), co-founder of the AI Now Institute, for the case that AI technology "cedes inordinate power" to a handful of corporations while significantly "capturing" academic research in the field. Estrin, who is the former CTO of Cisco, argues that the "hubris and determination of tech leaders to control society is threatening our individual, societal, and business autonomy."

fundamental skills; but she may wish to allow such use in an advanced course designed for those who have already mastered these skills.

- A professor might organize probing experiments that enable students to investigate model bias, perhaps preparing them to publish their results.

- An instructor teaching research at the graduate level may wish to allow students to use chatbots to improve grammar and syntax so long as they document that the actual research is their own.

- A humanities instructor who assigns writing and research to build critical thinking and sharpen engagement with course materials and themes may determine that use of chatbots poses a serious impediment to these objectives. He may therefore explain why the use of these tools is impermissible for assigned writing. However, he may simultaneously assign a research task in which students compare and contrast resources they found using search engines or library databases to those they found through chatbot use.

- An environmental science course may focus on the resource-intensive use of water and energy required to train and deploy chatbot systems without using these tools in class.

- A course in law or in graphic design may involve the study of "AI" copyright infringement across different companies and domains while inviting students to use these tools to audit models for research regarding intellectual property.

Of course, students will have their own views on the topic:

- Some may wish to opt out of using tools known to embed harmful stereotypes and/or subject users to surveillance and data collection.

- Some may seek to hone their ability to use "AI" in order to prepare themselves for the job market. They may find that a professor who assigns the probing of model bias and inaccuracies has prepared them to demonstrate their skills more deeply than a course that simply allows chatbot use for the generating of text.

- Some may recognize the virtue of probing models but have privacy concerns. They may request an alternative assignment that does not require them to sign up for a surveillant tool.

The good news is that all of these situations can effectively teach and enhance critical AI literacy, whether AI tools are used or not.

Several points are, however, worth emphasizing as you gear up for the new semester.

- **SYLLABUS:** Despite an already full workload complicated by the lingering effects of a global pandemic, we recommend that every instructor review their syllabus and their assignments in order to have the clearest possible policy regarding the use of AI (see the next section for advice).

- **ASSIGNMENTS:** In reviewing assignments, instructors may wish to implement changes in light of the fact that students may be tempted to use AI tools even if they are told not to do so. Simple response papers ("what did you think of this reading?") might work best in a classroom setting in handwritten fashion (or with wifi disabled on computers, phones, and tablets).[12] As an alternative for take-home assignments, consider project-based learning and/or scaffolded assignments (with some work taking place in the classroom) which might be suitable for such encouragement. Develop rubrics that emphasize critical thinking, problem-solving, and applied knowledge rather than memorization or summarization of content.

  For research papers, invite students to develop and research a topic they care about so as to encourage intrinsic motivation. Assign the use and citation of specific evidence, whether drawn from course materials or from independently researched sources. (Bot-generated text tends toward summary and generality with little or no quotation; when prompted to provide quotations, bots often deliver quoted material that is fabricated or incorrect).

  To further enhance motivation, provide students the opportunity to demonstrate their learning in different ways (*e.g.*, research paper, multimedia presentation, podcast, e-portfolio, etc.).

- **DETECTION**: The media discourse around student "cheating" is permeated by hype. Moreover, such discourse tends to portray college assignments as if they were task-specific labor disconnected from learning and the application of critical thinking. We believe that intrinsic motivation is one of the best ways to ensure a student's engagement with written work, research, and other forms of assessment.[13] A focus on cheating or

[12] Instructors should be prepared to accommodate students who have difficulty with handwriting, perhaps by opting for use of a wifi disabled device.

[13] See Lang (2013) for research that examines academic dishonesty and factors that "encourage" students to engage in this behavior. Lang also has three part series in the *Chronicle of Higher Education* on this research. (Part 1, Part 2, Part 3).

plagiarism, on the other hand, can undermine the experience for both instructors and students and change the relationship between them.

Please bear in mind that at present none of the systems being marketed to "detect" machine-generated text is reliable: false positives and false negatives are possible and even likely. Some of these tools evince biases against non-native English speakers.[14] The use of AI detection software, which is not FERPA-protected, may also violate students' privacy or intellectual property rights.[15] Instructors may wish to avoid such systems or at least to discount them as reliable evidence for violations of academic integrity.

Instructors who suspect the unauthorized use of an AI tool in their course should consider asking for a meeting to discuss the student's approach to completing the work and refer cautiously to the problematic content.

4. *Suggestions for Updating Your Syllabus So As to Clarify Course Policies and Learning Goals*

Whether an instructor wishes to build in the use of chatbots for certain assignments, allow students to experiment with them as they wish, or prohibit their use, we recommend clarifying these policies on syllabi and discussing them with students. Explain how you reached a decision that comports with the learning goals for the course. Consider discussing how chatbots work and the various problems described on this webpage (see potential resources below, including this recorded lecture by computational linguist Emily M. Bender). Members of Rutgers's advisory council are available with specific suggestions for teaching your students critical AI literacy in line with any of the below suggestions for syllabi.

I. For instructors who do not want students to use AI tools for their course

When specifying on one's syllabus that the use of chatbots and other AI tools is *not* permissible, instructors should be as clear as possible and may wish to refer to the Rutgers code of conduct, cited above, in doing so. Given that AI tools are (or may soon be) incorporated seamlessly into platforms such as Google Docs, grammar-checking tools such as Grammarly, and software suites such as Microsoft Office, a clear and specific statement is the best possible way to communicate with your students. In addition, you may wish to ask students to submit a statement of academic integrity along with their assignments.

For example,

---

[14] See Liang et al. (2023) for specific details about this study. Anecdotal accounts of outputs circulating on social media suggest that neurodivergent people may also be at risk for discriminatory assessment.

[15] In this context, it is worth noting that many of the digital tools students use voluntarily or according to instructor guidelines involve breaches of privacy and IP rights, including, at various points, Grammarly, Google Docs, and Zoom: on this changing landscape see, for example, Knowles (2023) and Merchant (2023).

In concert with Rutgers' code of conduct, which mandates "that all work submitted in a course, academic research, or other activity is the student's own and created without the aid of impermissible technologies, materials, or collaborations," this course has been designed to promote your learning, critical thinking, skills, and intellectual development without reliance on unauthorized technology including chatbots and other forms of "artificial intelligence" (AI). Although you may use search engines, spell-check, and simple grammar-check in crafting your assignments, you will be asked to submit your written work with the following statement. *"I certify that this assignment represents my own work. I have not used any unauthorized or unacknowledged assistance or sources in completing it including free or commercial systems or services offered on the internet or text generating systems embedded into software."* Please consult with your instructor if you have any questions about the permissible use of technology in this class.

Below is some alternative or additional language for syllabi which was developed at the University of Toronto.

- The use of generative AI tools or apps for assignments in this course, including tools like ChatGPT and other AI writing or coding assistants, is prohibited.
- The knowing use of generative AI tools, including ChatGPT and other AI writing and coding assistants, for the completion of, or to support the completion of, an examination, term test, assignment, or any other form of academic assessment, may be considered an academic offense in this course.
- Students may not copy or paraphrase from any generative artificial intelligence applications, including ChatGPT and other AI writing and coding assistants, for the purpose of completing assignments in this course.
- The use of generative artificial intelligence tools and apps is strictly prohibited in all course assignments unless explicitly stated otherwise by the instructor in this course. This includes ChatGPT and other AI writing and coding assistants. Use of generative AI in this course may be considered use of an unauthorized aid, which is a form of cheating.

II. For instructors who wish to permit use of AI tools in particular circumstances

When specifying on one's syllabus that the use of chatbots and other AI tools is *permissible in certain circumstances*, instructors should be as clear as possible and may wish to refer to the Rutgers code of conduct, cited above, in doing so. Bear in mind that students may be using these tools for different purposes in different classes so that it is important to be specific in describing the particular usages you allow or encourage. Given that AI tools are (or may soon be) incorporated seamlessly into platforms such as Google Docs, grammar-checking tools such as

Grammarly, and software suites such as Microsoft Office, a clear and specific statement that lays out permissible usages is the best possible way to communicate with your students.

For example, an instructor who does not want AI tools to be used in conjunction with written work but who wants to encourage students to do probing research on model content might consider the following statement:

In concert with Rutgers' code of conduct, which mandates "that all work submitted in a course, academic research, or other activity is the student's own and created without the aid of impermissible technologies, materials, or collaborations," this course has been designed to promote your learning, critical thinking, skills, and intellectual development without reliance on unauthorized technology including chatbots and other forms of "artificial intelligence" (AI). Although you may use search engines, spell-check, and simple grammar-check in crafting your assignments, you will be asked to submit your written work with the following statement. *"I certify that this assignment represents my own work. I have not used any unauthorized or unacknowledged assistance or sources in completing it including free or commercial systems or services offered on the internet or text generating systems embedded into software."*

A partial exception to this policy is an authorized exploration of model bias which we will conduct in Week X in order to build your learning on critical AI literacy.

Please consult with your instructor if you have any questions about the permissible use of technology in this class.

(As above, our recommendation is that any instructor assigning work that involves mandatory use of an AI tool consider developing an option for students who have data privacy or other concerns.)

Once again, we are sharing some alternative or additional language that was developed at the University of Toronto.

- Students may use artificial intelligence tools for creating an outline for an assignment, but the final submitted assignment must be original work produced by the individual student alone.
- Students may not use artificial intelligence tools for taking tests in this course, but students may use generative AI tools for other assignments.
- Students may use the following, and only these, generative artificial intelligence tools in completing their assignments for this course: …. No other generative AI technologies are allowed to be used for assessments in this course. If you have any question about the use of AI applications for course work, please speak with the instructor.

III. For instructors who wish to permit use of AI tools

When specifying on one's syllabus that the use of chatbots and other AI tools *is permissible (or encouraged)*, instructors should be as clear as possible about how this decision comports with the learning goals for their course and may wish to refer to the Rutgers code of conduct, cited above, in doing so. Instructors may also want to emphasize critical AI literacy including the importance of recognizing that current AI tools are subject to bias, misinformation, and "hallucinations" (as discussed above). Given that AI tools are (or may soon be) incorporated seamlessly into platforms such as Google Docs, grammar-checking tools such as Grammarly, and software suites such as Microsoft Office, a clear and specific statement about AI tools is the best possible way to communicate with your students.

For example, an instructor who encourages the use of AI tools in the course as a way to provide students with the opportunity to gain experience working with emerging technologies and to enhance their understanding of the topic might consider the following statement:

In concert with Rutgers' code of conduct, which mandates "that all work submitted in a course, academic research, or other activity is the student's own and created without the aid of impermissible technologies, materials, or collaborations," this course has been designed to help you develop knowledge and gain emerging skills that will be useful to you as workplace professionals. AI tools may be used as an aid in the creative process, but with the understanding that this should be accompanied by critical thinking and reflection. Students who choose to use these tools are responsible for any errors or omissions resulting from their use. They will also be required to provide as an appendix the prompts used, the generated output, and a thoughtful reflection on the outcomes. When appropriate, students may also be asked to consider the environmental and social costs of using the tool.

(As above, our recommendation is that any instructor assigning work that involves mandatory use of an AI tool consider developing an option for students who have data privacy or other concerns.)

Some instructors who permit use of AI tools for written assignments implement syllabus statements like these, developed at the University of Toronto.

- Students must submit, as an appendix with their assignments, any content produced by an artificial intelligence tool, and the prompt used to generate the content.

- Students may choose to use generative AI tools as they work through the assignments in this course; this use must be documented in an appendix for each assignment. The documentation should include what tool(s) were used, how they were used, and how the results from the AI were incorporated into the submitted work.

5. *A Short List of Resources You Might Wish to Read or to Share with Your Students*

This webpage already includes many resources that you might enjoy or might wish to share with or assign to your students. Here we provide a very short list. As this is a living document, we plan to continue to update it with additional resources as they become available.