AI & SOCIETY
https://doi.org/10.1007/s00146-022-01397-z

OPEN FORUM

How to cheat on your final paper: Assigning AI for student writing

Paul Fyfe[1]

Received: 13 September 2021 / Accepted: 4 February 2022

Abstract
This paper shares results from a pedagogical experiment that assigns undergraduates to "cheat" on a final class essay by requiring their use of text-generating AI software. For this assignment, students harvested content from an installation of GPT-2, then wove that content into their final essay. At the end, students offered a "revealed" version of the essay as well as their own reflections on the experiment. In this assignment, students were specifically asked to confront the oncoming availability of AI as a writing tool. What are the ethics of using AI this way? What counts as plagiarism? What are the conditions, if any, we should place on AI assistance for student writing? And how might working with AI change the way we think about writing, authenticity, and creativity? While students (and sometimes GPT-2) offered thoughtful reflections on these initial questions, actually composing with GPT-2 opened their perspectives more broadly on the ethics and practice of writing with AI. In this paper, I share how students experienced those issues, connect their insights to broader conversations in the humanities about writing and communication, and explain their relevance for the ethical use and evaluation of language models.

Keywords Language models · Plagiarism · AI literacy · Writing · Pedagogy · Ethics

> "The question then becomes, whose writing is this; who can take ownership? *The answer to this question is not easy to decide and seems to be more complicated than the question of whether it is true or false. I don't exactly have the right to claim ownership of it*, but I will argue that all writers borrow ideas and style from others." (undergraduate student with GPT-2 italicized)

This paper shares results from a pedagogical experiment that assigns undergraduates to "cheat" on a final class essay by requiring their use of text-generating AI software. For this assignment, students harvested content from an installation of GPT-2, then wove that content into their final essay. At the end, students offered a "revealed" version of the essay as well as their own reflections on the experiment. The epigraph above comes from a student's paper in which GPT-2 seems to join the conversation about the very ethical questions it has provoked. For their assignments, students were specifically asked to confront the oncoming availability of AI as a writing tool. What are the ethics of using AI this way? What counts as plagiarism? What are the conditions, if any, we should place on AI-assistance for student writing? And how might working with AI change the way we think about writing, authenticity, and creativity? While students (and sometimes GPT-2) offered thoughtful reflections on these initial questions, actually composing with GPT-2 opened their perspectives more broadly on the problems and practice of writing with AI. In this paper, I share how students experienced those issues, connect their insights to broader conversations in the humanities about writing and communication, and explain their relevance for the ethical use and evaluation of language models.

1 Research questions and frameworks

This teaching experiment invites students into an urgent conversation about the ethics of using AI in coursework. Especially for courses that involve writing, students and teachers must necessarily confront the problem of plagiarism with the wide availability of electronic sources and online essay mills. As will be discussed below, the concept of "plagiarism" itself needs more nuance, and it certainly gets blurrier in context of using an AI. But in its familiar form, especially as defined institutional settings like my own

✉ Paul Fyfe
paul.fyfe@ncsu.edu
http://go.ncsu.edu/pfyfe

[1] Department of English, NC State University, Raleigh, NC, USA

Published online: 10 March 2022

university, plagiarism appears as the uncredited, knowing, and sometimes wholesale adaptation of work that is not one's own.[1] For courses involving writing, that often manifests as text copied from online sources or acquired from vendors. Interestingly, artificial intelligence has upped the ante, disabling the usual strategies of detecting plagiarism by Googling phrases or checking papers against a known database (e.g. Turnitin). Some online resources now advertise AI to students to generate usable, unique text that is untraceable by current plagiarism detection software. For example, claiming in a tagline to be "Empowered by Artificial Intelligence," EssayBot.com promises itself as "your personal AI writing tool. With your essay title, EssayBot suggests most relevant contents. It paraphrases for you to erase plagiarism concerns." Marketing itself as a "bot" designed to outmaneuver plagiarism, the site feeds into concerns about the automation of writing and the erasure of human effort. In response, learning management systems and plagiarism detection software are now adapting AI tools of their own, locked in an arms race between crisply defined antagonists: systems to cheat artificially versus systems to insure original work. Staking out this battle line, one recent plagiarism detection product simply calls itself "Turnitin Originality."

However, this conflict presents a false binary, an overdrawn contrast for the sake of selling essays or student integrity insurance. These days, computer- and AI-assisted writing is already deeply embedded into practices that students already use. The question is, where should the lines be drawn, given the array of assistive digital writing technologies that many people now employ unquestioningly, including spellcheck, autocorrect, autocomplete, grammar suggestions, smart compose, and others? Asking students to "write with AI" can usefully provoke conversations not only about extreme examples of essay bots, but about everyday technologies, too. Within this spectrum of practices, what are the ethical thresholds? At what point, in what contexts, or with what technologies do we cross into cheating? Should that concept be redefined? The scholar Margaret Price has recommended that "the most constructive way to approach teaching on plagiarism is to invite students into a dialog about the subject, welcoming their perspectives on its complexities" (2002, 105). Those complexities have only increased with the ubiquity of AI. Offering a similar take on "How Teachers Can Prepare for AI-Based Writing," Jonathan Bailey (of Turnitin, note) suggests that teachers talk to students about these issues now, before "the inevitable day when students can push a button and the computer writes their paper" (Bailey 2020). That forecast seems a bit dramatic, but dramatic forecasts (as in science fiction about humans and AI) often help us reconsider the nuances of our own values, ethics, and relations to technology. In evolving forms, AI is the present and the near future of what faces students and educators.

Instead of a prescriptive approach to AI-assisted writing and plagiarism, assignments like this try something more proactive, allowing students to experience and then articulate for themselves the issues at stake. Furthermore, this approach moves students beyond antagonistic discussions of plagiarism to consider the potential uses of AI as a writer's tool. As Grant Otsuki argues, there's no putting the AI genie back in the bottle and little sense banning its use outright. Instead of teachers "pretending AI doesn't exist" or treating it merely as an antagonist, Otsuki encourages us to bring it into the classroom: "it might be time to train people to write *with* AI" (2020). What would that look like? While many writers are familiar with various assistive technologies, there are practically no examples for training writers with text-generating AI and language models like GPT-2.[2] Thus, the research questions from my experiment also explore students' strategies for AI-assisted composition. How might we write with these tools? What skills might writing with AI newly require? How does this challenge our assumptions about textual communication? And what potential risks and harms must we navigate? "Cheating" offers a starting point, but it opens onto much more complex and consequential topics, from human and nonhuman agency to the threats of disinformation and algorithmic bias. Thus, the assignment moves students from a familiar ethical context (is this plagiarism?) to new ethical questions (whose writing is this? what effects does it have?) and to new forms of writing practice (can I collaborate with AI?). The resulting insights can join a much broader conversation about AI literacy, including whether or how we might ethically and productively work with AI.[3]

Additionally, in this paper, I want to suggest how discussions of AI literacy can be enriched by ongoing scholarly conversations about plagiarism, authorship, and writing pedagogy. For example, composition and rhetoric scholars have been working thoughtfully on these issues for decades, having challenged the concept of plagiarism long before the ubiquity of the internet or machine learning. As

[1] Plagiarism appears as its own section of "Academic Misconduct" (8.4) in NC State's student conduct policy ("POL 11.35.01 – Code of Student Conduct" 2020).

[2] See Elkins and Chun for a report on an assignment oriented more toward literary writing (2020). See also the reflections by Lang et al. (2021).

[3] As defined by Long and Magerko, AI literacy includes "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (2020, 2). See also Ng et al. for a review of recent approaches to defining AI literacy (2021).

Price argues, "plagiarism is not stable" but situated differently "across historical time periods, across cultures, across workplaces, even across academic disciplines" (2002, 90). In our own time, educational institutions continue to define plagiarism in ways that idealize originality, a romanticized concept of a masculinized, independent, unified author which ignores all the rich dialogic messiness of language and the intertextuality of writing (Miller 2004; Flannery 1991). And emphasizing "originality" tends to punish the legitimate ways in which students come to learn a new discourse, often by imitation. For instance, in "patchwriting," students might swap or stitch their own words into a source text which, according to Rebecca Moore Howard, is part of a "healthy effort to gain membership in a new culture" (1992, 236). Howard reframes patchwriting as a "pedagogical opportunity, not a juridical problem" under the regime of plagiarism (1995, 789). Compositionists have also updated this pedagogical ethos for the digital age, suggesting how plagiarism anxieties tend to restrict the creative possibilities of networked multimedia and the "nature of writing in a remix culture" (Johnson-Eilola and Selber 2007, 381). When instructors strategically decriminalize plagiarism, they also open opportunities for creative and critical exploration. Along these lines, Johnson-Eilola and Selber propose that writing-as-remix activities "offer important new ways for thinking critically and productively about what it means to write, about what it means to read, and about what we value as texts" (2007, 377).

Throwing AI into the mix updates these questions in interesting ways—and helps to move discussions about language models beyond simplistic contrasts of human originality vs. machine imitation. That binary currently shapes so many responses to text-generating AI, whether in amazement about the coherent humanness of its outputs, or in sharp criticism of its failures to make sense or construct logical arguments (Branwen 2020; Marcus and Davis 2020; Seabrook 2019; Lea 2020). Yet such responses are entirely consistent with how AI—especially AGI, or artificial general intelligence—has historically been framed. In the long view, AI frequently appears as an imitation game or even as a practiced illusion (Standage and Stevenson 2018). Commentators still tend to fall into the Turing trap, as it were, evaluating any AI-generated language within a reductive framework of "bot or not" (Hua and Raley 2020). That reaction may show an anthropocentric bias, judging artificial intelligence only to the degree it resembles a human's. According to Miller, it also shows our deep "commitment to agency" that automation seems to challenge: in other words, it's not just whether synthetic writing sounds like a human; it's whether we can relinquish the humanistic concept of agency that has informed philosophy, politics, and pedagogy for a very long time (2007, 141).

Discussions of posthumanism are more open to the alternatives, and help signal how to approach writing with AI beyond the human/machine binary. Well before AI started to blend into our normal practices, Haraway declared that "writing is pre-eminently the technology of cyborgs" (2016, 57). Put differently, writing simply does not exist outside of the technologies used to produce it. And when technologies for writing change, so do its conceptual horizons [Perhaps writers change too, in a symbiotic process of "technogenesis" (Hayles 2012, 10)]. Writing with AI may force us into a heightened awareness of these posthuman dependencies, not only asking us to reexamine our definitions of writer, text, and reader, but to reevaluate our very identities within technological systems. What does that look like in practice? Certainly discussions of these issues are important, but active experimentation has a major role to play. As Casey Boyle argues, discussion alone preserves the split between humanistic minds and the technological objects they separately consider. Instead, Boyle recommends helping students to realize their embedded relationships to technological systems (2016, 540). A "posthuman practice" of writing, as Boyle calls it, would immerse students in mediated environments and stretch the ways they conceive of agents, text, networks, and communication. Perhaps most importantly for my argument, Boyle claims that such immersion offers a better place to create an ethics appropriate to our own complex technological systems. As he suggests, "ethics in a posthuman practice are not ideals imposed upon actions we ought do but are instead ongoing exercises whose aim is to compose new capacities for conducting ourselves within expanded media ecologies" (Boyle 2016, 549).

In such a spirit, this assignment places students within the media ecology of text-generating AI and asks them to articulate the ethics practiced therein. Plagiarism offers a familiar starting point for that exploration, and the varied ideas and critiques students propose—soon to be discussed—are interesting enough for continuing discussions of AI and writing. In a larger sense, students' experiments with GPT-2 also represent a different approach to AI literacy, including evaluating language models, their risks and harms, and the responsible practices of using (or not using) them.[4] As Minh Hua and Rita Raley have argued of the story-generating game AI Dungeon, its players become "necessary partners" with its underlying AI in a creative, critical, and functional relationship to it. Hua and Raley describe that relationship as a kind of "citizen NLP" and an example of "how humanists

[4] Long and Magerko recommend that "Researchers seeking to foster AI literacy may want to avoid misleading tactics like Turing deceptions and black-box algorithms" (2020, 8)—and I could not disagree more. Inviting students into this experience instead resembles how Meredith Broussard works with AI "to commit acts of investigative journalism" into its consequences (2018, 6).

might more meaningfully and synergistically contribute" to the assessment of language models (2020). While that's a reach goal for an undergraduate writing assignment, it underscores the value of their participation, even if—and especially if—those students do not hail from computational disciplines. In this spirit, I offer my students' experiments as test case of the kind of immersive, applied ethics that the humanities can bring.[5]

2 Materials and methods

This assignment was scaffolded into the final module of the honors course "Data and the Human" at NC State University (HON 202–006, Fall 2020).[6] The class enrolled 20 students from first-years to seniors with a variety of majors and concentrations. The course was designed as an interdisciplinary seminar, welcoming any motivated student, and requiring only the most basic technical abilities. Its goal was to cultivate what has been called "critical data literacy," combining a practical understanding of data and its transformations with a critical awareness of how data organizes knowledge and orchestrates power (Boyd and Crawford 2012; D'Ignazio and Klein 2020; Battista et al. 2020; Crawford 2021). The course carried students through three modules: data, privacy, and surveillance; definitions and critical dimensions of data; and finally machine learning and AI. Each module concluded with a hands-on experiment and accompanying report, in which students practiced with some of the technologies in question, and then reflected on what they experienced in context of the assigned readings and class discussions from that module. The NC State University Libraries generously provided students workshops for students to gain basic skills in all the tools the assignments required.

For the course's final module on AI, students had to write an essay that integrated the output from a text-generating language model into their own writing—but without revealing which was which. Styled as the "Professor Fyfe Turing Test," students were encouraged to try and make indistinguishable what essay text came from the AI or from themselves. I imposed no quotas for the amount of AI-generated text nor strictures about the ways it appeared. In the essay's final section, students composed exclusively in their own voices (i.e. without AI), reflecting on the following questions given in the assignment prompt:

> How easy or not was it to write this way? What worked or what didn't? How did the AI-generated content relate to your own? How did it affect what you might have thought about or written? Do you feel like you "cheated"? To what degree is this paper "your" writing? Do you expect a reader would notice GPT-2's text versus your own? Would you use this tool again, and in what circumstances? And, ultimately, what ideas about writing, AI, or humanness did the experiment test or change?

Finally, all papers included an Appendix in which students provided a "revealed" version of their essay with the AI-generated text highlighted. The assignment yielded 20 student essays of 1500 + words each. All students gave their permission to have their work considered and quoted in my own reports about the experiment.

Many of the assignment's questions were open ended, leading students into certain considerations while also letting them articulate their own insights. While evaluating these papers for our class, I captured responses to those guiding questions to include in the metrics and discussion to follow. I also compared those responses to students' academic disciplines and majors—which, as it turned out, did not factor at all into the patterns of responses. This exploratory project offers its findings as preliminary, encouraging more research perhaps with focused survey instruments or keying assignment questions and responses to emerging frameworks for AI literacy.

The language model students used to generate text was a medium-sized out-of-the-box installation of GPT-2, compassing 774 million language parameters. GPT-2 or "general purpose transformer" is an unsupervised language model trained on a large corpus from the web, released by OpenAI in 2019 (Radford et al. 2019).[7] It takes an initial "prompt" and attempts to predict, based on its learned parameters, what text is likely to follow. Scott Bailey, then a Digital Research and Scholarship Librarian at NC State, collaborated on the assignment design and created the interface for students.[8] We chose GPT-2 after reviewing what text-generation software, available at that time, might be most accessible to students.[9] We then created a user-friendly Python

[5] Such cross-disciplinary approaches might especially be warranted given the landscape of ethics training in CS programs (Raji et al. 2021).

[6] The course syllabus and module assignments are available open access in the MLA CORE repository: http://dx.doi.org/10.17613/0h18-5p41.

[7] For an extensive introduction to this class of machine learning software as well as its emerging uses and risks, see Bommasani et al. (2021).

[8] Scott has since moved to a position at the software company Sourcegraph.

[9] A quick demo of GPT-2 can be done online with the site "Write With Transformer" created by the company Hugging Face: https://transformer.huggingface.co/doc/distil-gpt2 In a subsequent version of this assignment, students used the open-source language model GPT-J which also makes a web-based demo available in a browser: https://6b.eleuther.ai/.

notebook in Google Colab which students copied into their own university-affiliated Google Drives to customize. With just a few clicks, students could easily activate the model and adjust a few settings, including the initial text prompt, the number of outputs, and the length of each output. Basically, they could enter their own text and have the model generate chunks of what it predicted came next. From that point on, "writing" the paper was up to them.

3 Results and discussion

At first glance, many students expected that this assignment would be easy, as if the time had arrived when they could "press a button and the computer writes their paper" (Bailey 2020). But students discovered that this was not easy at all. Most reported (87%) that it became far more complicated than just writing the paper themselves. When AI gets associated with automation or claims to save time and labor, applying AI to writing can appear like a shortcut. This may hold for generating text in shorter, formulaic genres like news spots or sports items, as we had discussed earlier in the module. But, as gets frequently observed, GPT-2 at least cannot sustain logically subordinated writing in longer genres like essays (Marcus and Davis 2020; Seabrook 2019; Elkins and Chun 2020).[10] Students quickly realized this for themselves, noting that "as it wasn't easy as I thought it would be," and "integrating artificially generated text into my writing was more of a curse than the blessing I thought it might be when this project was first explained," among similar reactions. All reported that GPT-2's outputs were difficult to control and often strayed off topic or into nonsense. They had to keep giving GPT-2 "a shove in the right direction," and then wrestle its outputs back into their contexts. At least without more extensive fine-tuning of the model, GPT-2 didn't often sound like them, and it certainly didn't follow where they wanted to lead.[11] These initial frustrations made it relatively easy to counter any student's received impressions of AI as an automatized writer.

Immediately confronted with its limitations, students also discovered GPT-2's capacity to fabricate information, including plausible-sounding false statements and even quotes from non-existent experts. For instance, one student noted that "sometimes the passage said the exact opposite of what was true, but the way it was worded seemed so professional and authentic I was almost convinced." The risks of such AI-generated misinformation have certainly been noted, including by OpenAI itself (Hern 2019). Even more striking than authoritative sounding text were quotes from non-existent authorities. Another student included this ironic quote from GPT-2 apparently about itself:

> "It's definitely a breakthrough, and I think it will have a significant impact on the field," said Professor Hervé Gagnon from the department of Design at the OpenAI and one of the researchers behind the project. "But I think it's not yet ready to be used by the average person," he added.

Of course, there is no such professor. Fabricated quotes like this shifted students' attention from plagiarism to the threats of "synthetic disinformation" and fake news (Kreps and McCain 2020). As one student reflected, "I think that the false information could be classified as cheating a lot more than the actual process itself." We had earlier read articles about the potential threats of GPT-2 in generating fake personae, expert testimony, and false reports (Metz and Blumenthal 2019). But it's a different matter when students become responsible for that material and whether or how to deploy it within their own essays, requiring a "shift in literacy practices" that accommodates algorithmic as well as human writing (Laquintano and Vee 2017).

Beyond its obviously fabricated personae, GPT-2 also raised questions for students about where or from whom it derived its statistical imitation of English. When GPT-2 did make sense, whose viewpoints did it express, and how could we even identify or cite them? Even if a student wanted to quote or credit GPT-2's output for a statement or an idea—in other words, even if they wanted to play fair by standard rules of plagiarism—how would they go about it? There are no citation styles for text generated by a language model, much less any disciplinary frameworks that would accept those contributions as valid. Conversely, whose viewpoints did GPT-2 exclude, and thus what paradigms could it never represent? Was GPT-2 a compromised source by default, untrustworthy because of its blind spots, and at worst a potential source of harm? Along these lines, several students reflected on the model's biases: as one put it, "[the assignment] showed me that programs like these do contain unjust prejudice from the data that they were trained on." As scholars have recently suggested of such language models, "large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations" (Bender et al. 2021, 1). We had prepared for these problems earlier in the module with assigned readings and discussion about algorithmic bias, which four students linked directly to their own experiments.

[10] Similar discussions can be found in the use of AI or automated systems to grade student writing (Anson 2006; Anson and Perelman 2017).

[11] We considered training the model on student's own writing for better results, but realized that would require a prohibitive and potentially invasive amount of training data from students: 50+MB their writing in plain text. We agreed a simpler approach would still achieve the assignment's goals.

Students rightly presented a lot of critique, perceiving their difference from GPT-2 in terms of viewpoints, goals, and writing style. Yet many students were also surprised by the degree to which GPT-2 seemed to resemble or even anticipate their own thoughts. For some, the writing partnership even worked nicely: "I was genuinely surprised with how well some of the content flowed with my personal writing and how it continued to sound like me." Several others likewise reported on the uncanniness of GPT-2's language imitation. Noting the "similarity between my writing and the predicted text," a student shared that "I felt like I was reading potential sentences that I would have written in another timeline, like I was glancing into multiple futures." In this case, AI unsettled the writer's orientation to narrative: an essay no longer unspooled along one single thread controlled by the author, but seemed almost fractal, as if multiplying the author and the possible pathways of their prose. It gave this student pause, at least, by unsettling the familiar dimensions of their writing and how they related to it.

A few students were startled that GPT-2 gave them substantive help with their essays. Occasionally, it did more than pump out plausible text: its outputs could contain useful ideas. A student revealed that "some of the things that the program came up with were not things I would have thought of on my own." In a related way, other students found that GPT-2 sometimes helped them articulate ideas still forming or that they struggled to express. As one explained: "those were the words that were on the 'tip of my tongue' and GPT-2 was able to articulate them for me in a better way." Echoing this statement, another student acknowledged that "GPT-2 aided me in formulating my thoughts, something which I notoriously struggle with." At a glance, these reports show students gratefully working through writer's block. Interestingly, these writers continue to claim the writing represents "my thoughts," just with mechanical help from GPT-2 in ordering them into grammatical sequence. But how can a student recognize the untrained outputs from a language model as their own ideas? In a sense, these students are "patchwriting," or borrowing and manipulating the structures of expert writing to join a discourse more credibly (Howard 1992, 233). According to Howard, patchwriting has real pedagogical value in helping students overcome self-consciousness about their writing or being outsiders. These students adapted AI outputs in a similar spirit: "the way the sentences were generated helped catalyze my thinking and writing process and helped me continue the paper. For me personally, I have all my ideas in my head but sometimes my brain just blanks out at how to write coherent and easy to read sentences and I end up getting frustrated." Fears about plagiarism can further cement this frustration, locking students into feeling disqualified because of their perceived deficiencies as writers.[12]

As students accommodated AI outputs into their essays, whether gratefully or begrudgingly, they all noticed different ways it impacted their writing. Some began to invent functions for GPT-2 as a writer's tool—ways that it might be used credibly and in earnest. However, even they acknowledged that results were somehow not their own. They faced a tradeoff: they could find more outputs matching their tone or style if they cared less about the content. For instance, "I was usually able to find sentences that flowed in a similar way to my writing even if it wasn't what I had in mind in terms of content." Other students took the opposite approach: changing their own writing style to better match GPT-2's outputs and "to create a smoother flow." One student realized that, to make this assignment work, "instead of desperately trying to make GPT-2 generate text to convey my ideas, I should just write my paper around whatever ideas GPT-2 conjured up." In these cases, GPT-2 was no longer articulating ideas for students, but dictating what they said and how they should write: "my own writing style adapted to more closely match the style of the generated responses rather than the other way around." That had negative consequences for some, as when a student judged that the "program decreased the quality of my writing. Since I was trying to closely match the writing style of the AI, I felt like I was not able to truly write like myself." Yet even that frustration may yield a net positive. Ironically, when encouraged to cheat, these students instead ended up reinforcing their own capacities as writers, the distinctiveness of their voices, and the tradeoffs if not outright sacrifices of using AI.

The eight students who concluded that they had indeed "cheated" on the assignment were committed to their own voice and pedagogical value of writing. Using GPT-2 seemed to short circuit the very intellectual goals of an essay: "I see writing papers as a far more expressive and individualistic manner of displaying comprehension … Writing is a skill that is able to be cultivated, but only through practice and understanding of your own identity." That student's reflection accords with some of the ways in-field experts also explain why neither writing nor evaluation can be mechanized. As Chris Anson summarizes, "The point of writing in a course is for students to explore and reflect on ideas through language, convey their own interpretations and informational discoveries to others, and in the process intersubjectively create purpose and meaning" (2006, 54). Asking students to write with AI may paradoxically help solidify these values, deepening the commitment of some

[12] This may especially apply in context of ESL and language diversity, as students and/or instructors disqualify their expression against standard written English.

writers to do without AI altogether. As a student concluded: "You'd think after years of writing papers I wouldn't think twice at letting a machine write one for me, but instead the opposite happened." Using GPT-2 only deepened their desire for control, as another student echoed: "Ironically, I feel like 'cheating' actually sabotaged my ability to write this lab report how I would have liked."

But other students were less sure they had really cheated, accepting GPT-2's contributions less as "plagiarizing a paper from a peer," and more as a kind of collaboration we have yet to define. In general, these students were more open to different configurations of authorship and writing. One student insightfully explained how this requires conceiving AI less antagonistically: "AI assistants were not meant to replace or impersonate humans but provide a bridge that connects our ideas with theirs—a hybrid." She moved beyond the binary of bot or not to understand AI as a collaboration between partners who are good at different things. This insight resembles how Ted Underwood has tried to reframe popular understandings of machine learning: "models will matter not because they're 'intelligent,' but because they allow us—their creators—to collaboratively explore a latent space of possibility" (2021).[13] Thus, this student did not mistake GPT-2 for a "peer" at all, but approached it instead as a tool for her own exploration. Such exploration became possible not only because of GPT-2, but because "the threat of punishment wasn't hovering over me," she said. This comment underscores two things. First, students can perceive creative and critical opportunities when allowed to experiment beyond the juridical framework of plagiarism. Inviting students to openly vs. clandestinely use AI may have shaped their responses to the assignment, but removing the threat of plagiarism opens participation in new ways and to writers of diverse backgrounds. Second, conventional definitions of plagiarism *already construct* AI as an intelligent peer, which is not only a mistake, but may foreclose on AI's more interesting possibilities, including what "hybrid" or collaborative configurations writers might discover.

Leah Henrickson has noted that "we do not know where computer-generated texts fit within our current conceptions of authorship and reading" (2021). In many ways, they don't fit at all, scrambling the ways we recognize and define how texts communicate. These ambiguities tended to interest students in the "not cheating" camp. As one wrote of their hybridized essay, "it doesn't feel like something I'd write but it also doesn't not feel like something I'd write." The vagueness of this statement is appropriate, as the demarcations of agency (what "I'd write") are no longer clear. The double negative—it's not not my writing—suggests our lack of vocabulary for recognizing ourselves in such an entangled relationship. Similarly, another student noted the strangeness of not recognizing the sources of their own essay after the fact: "when reading this essay, I often have to glance at the appendix to remind myself which phrases are not my original thoughts." They have not merely fooled themselves with a Turing test. Rather, this student implies how they're caught between a familiar framework for writing ("my original thoughts") and a hybrid one we have yet to build.

Several scholars have tried to articulate what this framework might be. Instead of using a "simple author-not author dichotomy" to judge computer-generated texts, Henrickson proposes that we place them along "a continuum from authorship to generatorship" (2019). This results in a more flexible concept that Henrickson calls "algorithmic authorship" (Henrickson 2019; 2021). David Rieder sees this less in terms of authorship than as the "domain of the digital rhetor," in which new configurations of digital media contribute to the rhetor's goals of persuasion.[14] Other scholars have proposed "co-creativity" and "synthetic literature" to describe writing with AI and the hybridity of computer-generated texts (Manjavacas et al. 2017). Notably, these arguments derive primarily from experiments in creative writing, new media art, and electronic literature rather than expository writing. That makes sense, given the relative openness of these studies to ambiguity and the sociotechnical interests of contemporary critics. Far less studied are how such hybrid frameworks might apply to argumentative prose, or how AI-generated materials might factor in scholarship on posthuman rhetoric and assemblage. Perhaps research and practice in new media arts point the way toward how language models might be used in good faith, used as "an instrument that extends, and fundamentally transforms, the human writing process" (Henrickson 2021).

Ten students overall answered that they had not cheated. Some of them hedged that while they didn't technically plagiarize, it still felt wrong. But about a quarter of the class had no need of such nuances. Intriguingly, they remained confident in their status as authors of their papers, and firmly defended their work as "not cheating" even within the conventional framework of plagiarism. They claimed to have done original, intellectual work—just not necessarily by *writing*. Instead, they explained their effort more in terms of assembly and editing. The AI could generate grammatical text, but until that text was sifted, organized, and stitched into an essay, the AI was not writing, per se. One student

[13] See also Reid on the "possibility space" of nonhuman rhetoric (2020).

[14] Personal conversation with the author. Rieder's own experiments in physical computing and new media composition are similarly interested in hybridity. He describes this as "everting" new media, or "infusing the real with aspects of the virtual."

reasoned that they hadn't cheated because "I ended up doing most of the work to help make the statements useful." Another agreed, making a fascinating comparison to explain their point: "I would say that I still wrote this paper. The difference is that I did not feel like a traditional writer. I felt more like Dr. Frankenstein, stitching together half sentences and incoherent AI words into something more cohesive." While the student did not pursue the analogy to *Frankenstein*, we might consider the provocative relation of suturing texts from GPT-2 to the surreptitious assembly of dead body parts, each resulting in an eloquent monster, each seeming to threaten the boundaries of the human, each asking hard questions about the relation of creature and creator. The student had played mad scientist in a text laboratory where ethics rules were temporarily suspended. And they realized that bringing such an essay to life required their substantive work. That realization—rather than their hideous progeny—made the experiment a success.

The intellectual labor of assembly includes the work of editing, which scholars have described as "not merely a mechanical procedure, but a complex and creative process" (Dragga and Gong 1989, 9). In their study of editing, Dragga and Gong detail the extensive "rhetorical repertoire" that editors must comprehend and employ in different situations (1989, 11). Within this repertoire, "arrangement" probably best describes what these students defended as their honest labor, having done all the work on the "organization of a text, the ordering of information according to appropriate cognitive patterns" (Dragga and Gong 1989, 12). Scholarship on writing-as-remix sees these thought processes activated in related ways. Johnson-Eilola and Selber propose the term "assemblage" and emphasize the critical, rhetorically-situated thinking it requires: "assemblages are texts built primarily and explicitly from existing texts to solve a writing or communication problem in a new context" (2007, 381).[15] In this light, while one student confessed they were "barely writing," they similarly defended their work as conceptual assembly: "it was more constructing an idea from the many voices." Ultimately, as another student concluded, "that involvement in my thought process turns the use of the generated text from cheating to not cheating … Doing all that work … makes me feel like I did more writing rather than less." Here, the word "writing" represents a far more complex, engaged activity than the act of putting down words, and using AI ironically required them to do more of it.

When writing with AI becomes more like the assembly and editing of texts, we change the role of author to something more like an editor, curator, or mediator. In these roles, students found themselves right in the middle of "a relationship between two entities who will attribute agency to each other"—authors and audiences, whose relationship seems profoundly unsettled by AI (Miller 2007, 149). Writing makes its meaning through the terms of that relationship, or what Henrickson calls the "hermeneutic contract" between authors and readers (2019). When new text technologies appear, those terms have to be renegotiated, eventually becoming normalized into the expectations for a given medium, to the point where we take them for granted. But no norms yet exist for writing with AI, no shared expectations for how it credibly makes meaning. This assignment asks students to renegotiate that hermeneutic contract, shifting writers and readers into a posthuman context where agency matters less. Put differently, dealing with the problems of expectations and norms makes student reflect upon the social protocols of an emergent text technology, rather than just seeing it as a tool. Thus, not only does this assignment aim at debunking AI as an automatized writer, it reframes writing with AI as a social negotiation in which students have a stake.

As Rieder puts it, this is the "age of digital persuasion." But in expository genres, writing with AI has a persuasion problem. How can readers be persuaded to accept its outputs? Where might AI make credible contributions to the writing process? What aspects of writing with AI can we agree to value, or in what contexts? We have already accepted many aspects of computational writing assistance into our norms: "the modern textual landscape," says Henrickson, is "permeated with varied modes of human–computer collaboration" (2019, 7). But text generation continues to be a sticky problem, and consensus is not coming soon. My students realized that we were nowhere close to push-button solutions to writing essays automatically, but that writing with AI has emerged as an important issue for us all to grapple with now, to reject with reason, and/or to reimagine with specific collaborative practices.

4 Closing

Collectively, my class did not conclude one way or another that writing with AI was tantamount to plagiarism. They split down the middle on the question, with all sorts of qualifications about their votes. But this assignment was never meant to settle a debate. Instead, it immersed students within the debate to broaden and deepen their perspectives, and even to raise other questions we had not yet collectively discussed. As one student perceptively judged: "Overall, the experiment was able to test multiple ideas … It was interesting to experience these ideas after having read about them." The "hands-on" experience let students explore the arguments for themselves, as well as test out new configurations

[15] For an extensive introduction to assemblage in writing studies, as well as current examples in pedagogical practice, see Yancey and McElroy (2017).

of writing beyond the potentially limiting concept of humans vs. AI. Furthermore, the assignment proved useful not only in reckoning with text-generating AI, but in having students reflect on the values of their own expression. The assignment's framework of AI-powered plagiarism led students to assume that they would somehow avoid the intellectual work of writing. But the opposite happened, reinforcing or expanding students' definition of what that work encompassed. As the author Robin Sloan reports of his own experiments writing with machines, "the goal is not to make writing 'easier'; it's to make it harder" (2016). In other words, such experiments refuse to let us take our ideas about writing for granted, pushing us to rearticulate or re-envision them. And, happily, even though this required more work and some occasional aggravation, students were engaged in ways surpassing my expectations.

Whether or not writing with AI becomes cheating is a much more complicated question than software purveyors and institutional plagiarism policies would have us believe. And humanities scholars can play an important role in helping explain and navigate those complexities. For starters, we should reframe that originating question in a few different ways. How do institutionalized definitions of plagiarism block the possibilities of writing with computational assistance or even partnership? What kinds of creative or critical frameworks do plagiarism policies impose or preclude, especially compared to what we really value for students? How are these values supported (or not) by the learning activities that AI seems likely to impact (like writing essays)? How might we modify these activities or emphasize different outcomes? How do plagiarism regimes or AI platforms differently affect populations of students? Are there new modes of creativity, critical reasoning, rhetoric, assemblage, and expression that computational assistance helps us identify? How might AI-assisted writing exercises align with frameworks for teaching AI literacy?[16] What are the risks and harms of these technologies that we need to factor in? And how can we bring different disciplinary perspectives on these questions to bear upon the development and implementation of AI?

This essay has recontextualized AI-powered plagiarism with students' reflections alongside examples from scholarship in writing studies, rhetoric, book history, media theory, HCI, and science and technology studies. All of these fields—in their work past and present—continue to offer vital perspectives on whether or how writing with AI might be embedded in society. Assignments like this may help amplify for students, too, a call for their participation in the discourse. My course included students from various disciplines and, as an English professor, I especially wanted to embolden students specializing in fields other than computer science.[17] But CS students, too, often lack programmatic opportunities for creative and critical experimentation. The most glowing review of this technically simplistic assignment came from a student in computer engineering. Another CS student chose to pursue internships in AI and ethics as a result. As Hua and Raley argue, we all have a stake "in training, evaluating, and collaborating with the autonomous systems that will continue to speak and write on our behalf" (2020). Embedding students within these debates helps show them not only the important issues at stake, but invites them into the evolving ethical project of dealing with AI in our world.

Acknowledgements My thanks to the students in the HON 202 seminar "Data and the Human" for their enthusiastic participation in this experiment. Many thanks also to NC State University colleagues including Scott Bailey, Zachary Beare, Chris Anson, Helen Burgess, and David Rieder for sharing their perspectives and recommendations. Thanks also to peer reviewers for constructive suggestions that improved this article.

Author contributions Not applicable.

Funding Not applicable.

Availability of data and material The course syllabus and module assignments are available open access in the MLA CORE repository: https://doi.org/10.17613/0h18-5p41.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Additional declarations for articles in life science journals that report the results of studies involving humans and/or animals Not applicable.

Ethics approval Federal regulations allow specific categories of human subjects research to be exempt from continuing IRB review [45 CFR 46.104(d)]. Exemption Category 1 applies to research conducted in established or commonly accepted educational settings involving normal educational practices, including research on classroom instructional strategies.

Consent to participate All enrolled students completed the major assignments as course requirements.

Consent for publication All students gave their emailed consent to have their essays included and quoted in my report. Quotes from student papers have been anonymized in this article.

[16] For specific examples from those frameworks, see Long and Magerko (2020), Ng et al. (2021).

[17] As Lauren Goodlad complains, "while there is increasing talk of making AI 'ethical,' 'democratic,' and 'human-centered,' scholars in the humanities seldom shape these discussions" (Goodlad and Dimock 2021, 317).

References

Anson CM (2006) Can't touch this: reflections on the servitude of computers as readers. In: Patricia FE, Richard H (eds) Machine scoring of student essays: truth and consequences. Utah State University Press, Logan, pp 38–56

Anson CM, Perelman L (2017) Machines can evaluate writing well. In: Ball CE, Loewe DM (eds) Bad ideas about writing. Digital Publishing Institute/West Virginia University Libraries, Morgantown, pp 278–286

Bailey J (2020) How teachers can prepare for AI-based writing. Turnitin. https://www.turnitin.com/blog/how-teachers-can-prepare-for-ai-based-writing. May 21, 2020.

Battista A, Katherine B, Marybeth M (2020) Data literacy in media studies: strategies for collaborative teaching of critical data analysis and visualization. J Interact Technol Pedag. https://jitp.commons.gc.cuny.edu/data-literacy-in-media-studies-strategies-for-collaborative-teaching-of-critical-data-analysis-and-visualization/.

Bender EM, Timnit G, Amy M-M, Schmargaret S (2021) On the dangers of stochastic parrots: can language models be too big?" In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 1–14. Virtual Event, Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bommasani R, Drew AH, Ehsan A, Russ A, Simran A, Sydney VA, Michael SB, et al (2021) On the opportunities and risks of foundation models. http://arxiv.org/abs/2108.07258.

Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inform Commun Soc 15(5):662–679. https://doi.org/10.1080/1369118X.2012.678878

Boyle C (2016) Writing and rhetoric and/as posthuman practice. Coll Engl 78(6):532–554

Branwen G (2020) GPT-3 creative fiction. Gwern.Net. 2020. https://www.gwern.net/GPT-3.

Broussard M (2018) Artificial unintelligence: how computers misunderstand the world. MIT Press, Cambridge

Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven

D'Ignazio C, Klein L (2020) Data feminism. MIT Press, Cambridge

Dragga S, Gong G (1989) Editing: the design of rhetoric. Routledge, London

Elkins K, Jon C (2020) Can GPT-3 pass a Writer's turing test? J Cult Anal. https://doi.org/10.22148/001c.17212

Flannery KT (1991) Composing and the question of agency. Coll Engl 53(6):701–713. https://doi.org/10.2307/377895

Goodlad LME, Dimock WC (2021) AI and the human. PMLA 136(2):317–319. https://doi.org/10.1632/S0030812921000079

Haraway DJ (2016) A cyborg manifesto: science, technology, and socialist-feminism in the late twentieth century. In: Manifestly Haraway. University of Minnesota Presshttps://doi.org/10.5749/minnesota/9780816650477.003.0001

Hayles NK (2012) How we think: digital media and contemporary technogenesis. The University of Chicago Press, Chicago London

Henrickson L (2019) Towards a new sociology of the text: the hermeneutics of algorithmic authorship. PhD Dissertation, Loughborough University.

Henrickson L (2021) Reading computer-generated texts. Camb Univ Press. https://doi.org/10.1017/9781108906463

Hern A (2019) New AI fake text generator may be too dangerous to release, say creators. The Guardian. https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction. February 14, 2019

Howard RM (1992) A plagiarism pentimento. J Teach Writ 11(2):233–245

Howard RM (1995) Plagiarisms, authorships, and the academic death penalty. Coll Engl 57(7):788–806. https://doi.org/10.2307/378403

Hua M, Rita R (2020) Playing with unicorns: AI dungeon and citizen NLP. Digit Hum Quart 14(4). http://www.digitalhumanities.org/dhq/vol/14/4/000533/000533.html.

Johnson-Eilola J, Selber SA (2007) Plagiarism, originality, assemblage. Comput Compos 24(4):375–403. https://doi.org/10.1016/j.compcom.2007.08.003

Kreps S, Miles M (2020) Not your father's bots. https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots. April 16, 2020.

Lang A, Quinn D, Annie KL (2021) The ghost in Anouk's Laptop. The Data-Sitters Club (blog). https://datasittersclub.github.io/site/dsc9. February 17, 2021

Laquintano T, Annette V (2017) How automated writing systems affect the circulation of political information online. Lit Compos Stud 5(2):43–62. https://doi.org/10.21623/1.5.2.4

Lea R (2020) If a novel was good, would you care if it was created by artificial intelligence?. The Guardian. http://www.theguardian.com/commentisfree/2020/jan/27/artificial-intelligence-computer-novels-fiction-write-books. January 27, 2020

Long D, Brian M (2020) What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–16. CHI '20. Association for Computing Machinery, New York https://doi.org/10.1145/3313831.3376727.

Manjavacas E, Folgert K, Ben B, Mike K (2017) Synthetic literature: writing science fiction in a co-creative process. In: Proceedings of the Workshop on Computational Creativity in natural Language Generation (CC-NLG 2017), 29–37. Association for Computational Linguistics, Santiago de Compostela. https://doi.org/10.18653/v1/W17-3904.

Marcus G, Ernest D (2020) GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about—MIT technology review. In: MIT Technology Review (blog). https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. August 22, 2020

Metz C, Scott B (2019) How AI could be weaponized to spread disinformation. The New York Times, June 7, 2019.

Miller S (2004) Rescuing the subject : a critical introduction to rhetoric and the writer. In: Paperback (ed). Southern Illinois University Press, Carbondale

Miller CR (2007) What can automation tell us about agency? Rhetor Soc Q 37(2):137–157. https://doi.org/10.1080/02773940601021197

Ng DT, Kit JK, Leung L, Chu KWS, Qiao MS (2021) AI literacy: definition, teaching, evaluation and ethical issues. Proc Assoc Inform Sci Technol 58(1):504–509. https://doi.org/10.1002/pra2.487

Otsuki GJ (2020) OK computer: to prevent students cheating with AI text-generators, we should bring them into the classroom. In: The Conversation (blog). https://theconversation.com/ok-computer-to-prevent-students-cheating-with-ai-text-generators-we-should-bring-them-into-the-classroom-129905. January 23, 2020.

POL 11.35.01—code of student conduct (2020) NC State University. 2020. https://policies.ncsu.edu/policy/pol-11-35-01/.

Price M (2002) Beyond 'gotcha!': situating plagiarism in policy and pedagogy. Coll Compos Commun 54(1):88–115. https://doi.org/10.2307/1512103

Radford A, Jeffrey W, Dario A, Daniela A, Jack C, Miles B, Ilya S (2019) Better language models and their implications. OpenAI. https://openai.com/blog/better-language-models/. February 14, 2019.

Raji ID, Morgan KS, Razvan A (2021) You can't sit with us: exclusionary pedagogy in AI ethics education. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,

515–25. FAccT '21. Association for Computing Machinery, New York. https://doi.org/10.1145/3442188.3445914.
Reid A (2020) Synthetic speech and negotiation: AI's nonhuman rhetoric. Enculturation. http://enculturation.net/synthetic_speech_and_negotiation.
Seabrook J (2019) Can a machine learn to write for the new yorker? https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker. October 7, 2019.
Sloan R (2016) Writing with the Machine. Robin Sloan (blog). 2016. https://www.robinsloan.com/notes/writing-with-the-machine/.
Standage T, Seth S (2018) The box that AI lives. In: The Secret History of the Future. http://www.slate.com/articles/podcasts/secret_history_of_the_future/2018/09/a_200_year_old_chess_playing_robot_explains_the_internet.html. Accessed September 28, 2018.
Underwood T (2021) Science fiction hasn't prepared us to imagine machine learning. In: The Stone and the Shell (blog). https://tedunderwood.com/2021/02/02/why-sf-hasnt-prepared-us-to-imagine-machine-learning/. February 2, 2021
Yancey KB, Stephen M (eds) (2017) Assembling composition. In: CCCC studies in writing and rhetoric. In: Conference on College Composition and Communication of the National Council of Teachers of English, Urbana