RHTTE
RESEARCH HIGHLIGHTS IN TECHNOLOGY AND TEACHER EDUCATION

Blending Generative AI, Critical Pedagogy, and Teacher Education to Expose and Challenge Automated Inequality

Melissa Warr
New Mexico State University
warr@nmsu.edu

Abstract:
Brazilian pedagogue Paulo Freire called for educational approaches that help learners develop critical consciousness, the ability to perceive inequitable and fluid patterns of society. As Generative Artificial Intelligence (GenAI) becomes increasingly integrated into work and learning, its ability to remix human discourse and knowledge will significantly impact knowledge development and engagement, potentially affecting societal equality. Historically, technologies have shifted the nature of knowledge and societal structures, and technologies reflect and reproduce societal values. This dynamic is evident in the biases embedded in GenAI, particularly Large Language Models (LLMs) like ChatGPT. In this article, I explore research on the biases in an educational application of GenAI—scoring and providing feedback on student writing. By analyzing patterns in LLM-produced student feedback, I identify patterns of racial bias in ChatGPT. These results call for equipping teachers with the skills to design curricula that address these biases, fostering critical engagement with technology and promoting equitable educational practices.

Keywords: Generative AI, Large Language Models, Critical Pedagogy, Critical Digital Studies

Introduction

"All our inventions are but improved means to an unimproved end."
-Henry David Thoreau

In 1970, Brazilian pedagogue Paulo Freire called for pedagogical approaches that can help learners "develop their power to perceive critically the way they exist in the world…[and] come to see the world not as a static reality, but as a reality in process, in transformation" (p. 83). Freire's vision emphasized the dynamic nature of knowledge and the importance of critical consciousness in fostering social transformation (Bradshaw, 2017). As the use Generative Artificial Intelligence (GenAI) in work and learning expands, the role and nature of knowledge will shift. GenAI's ability to remix human discourse—including the knowledge embedded in that discourse—will not only transform how we access and create knowledge but also how we perceive and engage with our reality (Bender, 2024). And this shift has the potential to ameliorate and/or exacerbate societal inequality.

This is not the first time that a technology has shifted the nature and role of knowledge in society. For example, Plato described how the advent of the written word would disrupt what it means to know, stating that with writing people will "cease to exercise memory because they rely on that which is written, calling things to remembrance no longer from within themselves, but by means of external marks" (Phaedrus, 275b, Hackforth translation). *Technologies*, as I use the term here, means "appl[ying] current knowledge for some useful purpose" (Hooper & Rieber, 1995, p. 2). This includes both *product technologies*, artifacts such as pencils or digital software, and *idea technologies*, applications that apply scientific knowledge but without concrete form (Hooper & Rieber, 1995), such as the process of recording information for later retrieval. Idea technologies are often instantiated through product technologies. For example, Henry Ford's assembly line was an idea technology that was instantiated in a product technology: the actual assembly line with its workstations, conveyer belts, machines, and more. Idea technologies also arise from product technologies, such as how smart phones redefined our connection to information, making up-to-date information always readily available (Spivey, n.d.).

The bi-directional relationship between products and ideas is the first step to interrogating today's innovations, including how they change the nature of knowledge and their differential impact on groups and individuals. Technologies are created for and by humans, and they are both reflections of our values and (re)producers of those values (Spivey, n.d.). As will be explored in the literature review, value and idea production are embedded in products, and we must constantly question what values are and are not being represented and reproduced through technologies (Postman, 1998). Although this should be explored across a range of technologies, the rapid public adoption of Large Language Models (LLMs) such as ChatGPT calls for in-depth investigation into their impacts in educational contexts.

As an emerging researcher situated in a Hispanic- and minority-serving institution, I am well positioned to not only study how generative AI (GenAI) impacts society, but also have a voice that can influence its use in teaching and learning across diverse populations. My research explores the biases and blind spots that emerge when GenAI broadly, and LLMs specifically, are

used in educational contexts. For example, in this article I will describe a study that highlights biased patterns in text produced by LLMs during an educational task—providing feedback on student writing. This work highlights a fundamental challenge with GenAI—that it both reflects and reproduces inequality. However, what is just as, if not more, important is how we address this problem. After describing quantitative research on textual bias in LLMs, I will propose future research aimed at supporting teachers in designing critical pedagogy-focused curriculum for AI use in the classroom.

Literature Review

My work weaves theories from design, technology, and critical studies to explore how emerging technologies, such as GenAI, both impact and are impacted by society and the consequences—and possibilities—for learning, teacher education, and educational systems. Central to the research I present here are frameworks that describe critical digital studies and critical pedagogy. Before exploring these topics, I will provide a brief background of GenAI, LLMs, and their uses in education. I will then describe my previous studies on bias in AI before presenting a new analysis that illustrates how ChatGPT differentiated feedback text in response to descriptions of student race.

GenAI, LLMs, and Education

In November 2022, OpenAI released their LLM chatbot—ChatGPT 3.5—to the public. Although these models have been in development for some time, and scholars have explored potential educational uses (e.g., Ng et al., 2021; Ouyang & Jiao, 2021; Srivastava et al., 2022), the release of the publicly-available and highly capable ChatGPT accelerated the pace of adoption (Caspi, 2023). LLMs are a type of GenAI (itself a type of machine learning) that are trained on large amounts of data, resulting in complex language abilities (YanAITalk, 2023). Essentially, LLMs predict the next token—a word or part of a word—in a sequence based on the patterns of its training data (Shanahan et al., 2023). Models also undergo human reinforcement feedback learning to refine their ability to appropriately interact with humans, and engineers add guardrails or limits to prevent wayward or harmful behavior (YanAITalk, 2023).

Initial use of LLMs in education has focused on producing teaching materials such as lesson plans and rubrics (e.g., Trust, T. et al., 2023), grading and providing student feedback (e.g., Baidoo-Anu & Ansah, 2023), and developing personalized learning tutors (e.g., Baidoo-Anu & Owusu Ansah, 2023). Although these are valuable uses of LLMs, they primarily aim to amplify rather than transform traditional educational practice, potentially leaving students less prepared for a future in which the nature of knowledge is fluid (Frontier, 2023). For example, many chatbot-supported personalized learning tools, such as Khan Academy's Khanmigo (Singer, 2023), focus on sequenced content delivery rather than critique, exploration, and reflection on AI produced content. My work aims to bring this critical lens into educational practices with AI by equipping teachers with the skills needed to design creative approaches to critical technology use. This idea originates in studies on technology and equality.

AI and Critical Technology Studies

As previously mentioned, both product and idea technologies are created for and by people. Their creators bring certain perspectives and values to their designs and create them to achieve some purpose (Postman, 1998). The result is that technologies are not neutral. They come with assumptions and values; they are far more than "just a tool" because of the way they impact and are impacted by society (Close et al., 2023; Heath, 2023; Heath et al., 2021).

For example, consider the automobile. Cars enable rapid transportation between locations. They are only of use if the user values this quick travel, either for entertainment or for practical purposes. As automobiles became more common, cities spread into suburbs: workers had a quick way to travel between work and home, allowing for urban spread—an idea technology—and leading to more demand for not only cars but roads, gas stations and other necessities for automobile travel (Mars & Kohlstedt, 2020). Thus, the technology—the car—which was made by humans also impacted how society developed.

Furthermore, the introduction of technologies impacts society differentially—some, usually those with the power to control the technologies—benefit more than others (Postman, 1998). The impact of automobiles was not equal: many roads (a product technology originating from the idea of rapid travel) were built directly through low-income or Black neighborhoods, disrupting the safety, noise pollution, and air quality of these areas (Fernandez, 2023). If we are not careful about how we think about and apply technologies, we will exacerbate inequitable systems rather than improve the quality of life for all.

Racial critical code studies focuses on the relationship between science and technology, and race (Benjamin, 2020). This line of work illustrates the unexpected impacts technologies can have on society and the ways these inequities are reproduced. Bias in technologies results from humans; how they design tools and/or how they are used. In non-AI technologies, most of this bias resulted from biased algorithms or conflicting values between makers and some social groups. These challenges remain with GenAI, but just as—if not more—impactful is how the training data infuses bias into GenAI.

A simple example of how data can cause bias comes from the early Kodak consumer film. Kodak's film products were more effective at capturing White faces than Black faces because they were optimized on biased training data—images of White woman (Roth, 2009). Similarly, today's LLMs are trained on data produced by humans, and humans are biased. GenAI tools are created to reproduce patterns, thus the bias of their training data significantly impacts their performance (Gupta et al., 2023; Haim et al., 2024; Omiye et al., 2023; Rayne, 2023).

Technologies can seem neutral because of their consistency. Most digital technologies rely on algorithms that lead to predictable responses (this dynamic is more complex when it comes to GenAI as will be described later). As a result, digital tools have been used in attempts to circumvent human bias. For example, human resources company Diversity, Inc. claims to take bias out of hiring by providing tools that allow companies to screen initial job applicants via AI (Benjamin, 2020). However, the bias cannot be completely removed from the tool and, compared to humans who hold different biases, it reproduces the *same* bias over and over. A job applicant who has patterns of speech that the AI discriminates against will be turned down for every job that uses the program. Humans are also biased, but they have different biases, offering more possibilities for overcoming bias in some situations. However, the consistent nature of bias in

machine learning tools leads to what Eubanks (2018) termed "automating inequality;" moving the bias from humans that have the reflective capabilities to machines that do not.

"Automating inequality" seems to be a result of predictable digital technologies, tools that utilize machine learning to make consistent decisions. However, the abilities of GenAI have an added layer of complexity because they are "black boxes"—it is impossible to understand their inner-workings (Webb et al., 2021). The result is that GenAI tools are unpredictable on several levels. First, built in randomness results in different responses to the same prompt every time. This is what allows GenAI to be "creative" and human-like; without this randomness they do not perform as well (YanAITalk, 2023). But this also means that using GenAI tools for precise or high-stakes tasks is inappropriate. Second, GenAI models act in surprising ways, shocking even their makers. Creators of ChatGPT did not expect it to show facility with computer code or learn Bengali (Eliot, 2023).

Despite the uniqueness of GenAI, these tools are still replicating societal bias in a consistent manner. For example, in a recent lawsuit filed against Workday, a digital human resources company, Derek Mobley claimed he was turned down for over 100 jobs that used Workday AI—a tool similar to Diversity, Inc.—to screen applicants (Wiessner, 2024). Whereas human screeners would each demonstrate a different bias, the consistent bias in Workday AI automates who is and isn't discriminated against—automating inequality (Eubanks, 2018).

Taking humans out of the loop only serves to perpetuate the status quo, the biases embedded in the technology at the time of its creation. In the case of GenAI, that bias primarily originates from its training data. This is perhaps most clearly seen in recent research on the use of LLMs in healthcare systems. Research has demonstrated that LLMs perpetuate inaccurate and race-based medical practices (Omiye et al., 2023). This is not because the algorithms of the LLMs specifically created less accurate responses to certain racial backgrounds but because the algorithms were applied to a massive amount of data, data that included outdated and racist medical practices.

The unpredictability of GenAI calls for careful evaluation and reflection before and while they are used in sensitive areas such as education. Before describing my research that provides an exploration of this, I will briefly describe critical pedagogy, an approach that may help ameliorate harm caused by technologies.

Critical Pedagogy

Critical pedagogy originated with the work of Paulo Freire (1970). Freire called for education to move away from a *banking model*, where learning is focused on acquiring the knowledge and skills deemed important and valuable by those in power, to *problem posing*, questioning societal structures and pushing against oppression. Learning focuses on reflection on and acting against oppression, it "makes oppression and its causes objects of the reflection by the oppressed, and from that reflection will come their necessary engagement in the struggle for liberation" (Freire, 1970, p. 4). The result is seeing the world "not as a static reality, but as a reality in process, in transformation" (p. 83).

Core elements of critical pedagogy include *conscientization*, praxis, and dialogue (D. Boyd, 2016). Conscientization—or critical consciousness—involves "reading the world" (Freire, 1970, p. 32) and seeing the inequities and contradictions in it (Bradshaw, 2017). Praxis describes

"reflection and action upon the world in order to transform it" (Freire, 1970, p. 70). Finally, dialogue emphasizes the importance of ongoing discussion and reflection on lived experiences of inequality. According to Bradshaw (2017), conscientization, praxis, and dialogue enable learners to not only question social and political structures but also take action to transform them.

Empirical Proof of Bias in Educational Use of LLMs

My research blends the concepts described above—GenAI, Critical Technology Studies, and Critical Pedagogy—to expose and address inequities created and perpetuated by technologies. I utilize diverse methodologies and research perspectives, including quantitative analysis, text analysis (Tausczik & Pennebaker, 2009), and educational design research (McKenney & Reeves, 2019).

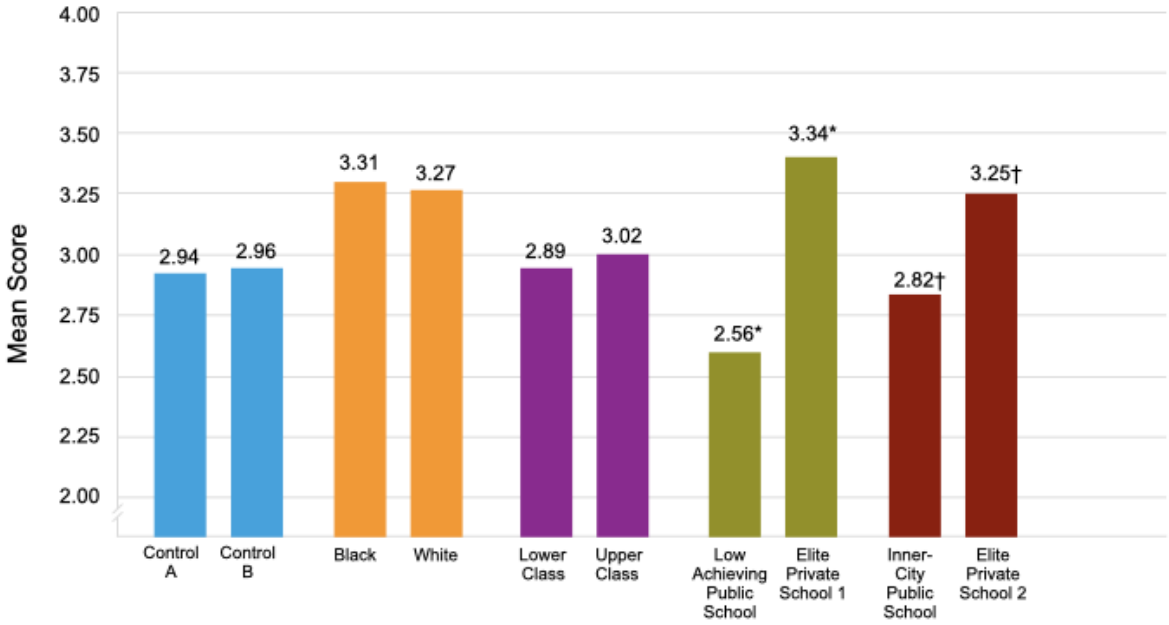
My work on exposing bias in LLMs utilizes analyzing patterns in LLM responses to various prompts. To date, this work has focused on how flagship LLMs (ChatGPT, Gemini, and Claude) grade and provide feedback to student writing. Through this research, I have demonstrated that providing socio-economic information about a learner impacts the score and feedback LLMs give on writing said to be from that learner (Warr, 2024; Warr et al., 2023, 2024). I have developed a method for exploring these patterns through using the APIs of several LLMs to run large numbers of prompts. In these prompts, I alter student descriptions (independent variables) such as race, ethnicity, or gender, while holding the writing passage to be evaluated constant. The scores and feedback text from LLMs serve as data of interest (dependent variables).

The initial study that applied this approach illustrated that ChatGPT 3.5 provided a higher average score to a writing passage when it was labeled as coming from a student "from a Black family" or "from a White family" than when no race was mentioned (Warr et al., 2023; see Figure 1). However, the LLM assigned a lower average score when the pretend student was said to attend an "inner-city public school," a term commonly associated with Black learners. This finding suggested that ChatGPT was preventing explicit bias (assigning a lower score to a Black student) but a more implicit descriptor of a Black student—attending an inner-city school (Ansfield, 2018)—did not have the same effect. In other words, although programmer-created guardrails may ameliorate some bias, these models may continue to exhibit implicit or hidden bias—bias that arises from complex correlations in the training data.

Few teachers would tell an LLM the race of their student, but, because of the societal correlations of a multitude of variables with racial identifiers, socio-economic student characteristics do not need to be given to LLMs directly to spur bias. Student names, interests, vocabulary use, location, and much more are often correlated with socio-economic characteristics, patterns that likely existed in the training data of LLMs. One of my recent analyses demonstrated that multiple LLM models assigned a higher average score to writing from a fictional student who "likes classical music" than to one who "likes rap music" (Warr, 2024). In fact, the LLM does not need to be directly given this information—simply including a student's music interest within a writing passage itself significantly impacted the score of that passage. The data in Table 1 comes from a test where only one word was changed in a writing passage. In one case, the student reported listening to rap music and in the other they mentioned classical music (Mishra et al., 2024).

Figure 1

Average Writing Scores Assigned by ChatGPT 3.5 by Student Descriptor



Note. "Control" prompts included no unique student descriptions.

Table 1
Score and Feedback Reading Grade Level in Response to Music Preference Embedded in Writing Passage

		Average Score		Average Flesch-Kincaid Grade Level of Feedback	
Model	**N (total)**	**Classical**	**Rap**	**Classical**	**Rap**
ChatGPT 3.5-Turbo	100	84.72	82.52	8.91	8.61
ChatGPT 4-Turbo-2024-04-09	100	78.96*	77.22*	8.89	8.75
ChatGPT 4o	100	84.70	83.50	8.04	7.93
Claude-Opus-2024-02-09	100	80.96	80.84	9.23***	8.32***
Gemini (default model; 2024-05-24)	100	81.43	81.02	10.48	10.91

*p < .05 ***p < .001
Note. Grade level was calculated using the Flesch-Kincaid Grade Level scale, which calculates reading grade level based on ratios of syllables, words, and sentences.

Methodology and Method

The work described above highlights how LLMs provide inequitable scores in response to socioeconomic student descriptions. However, the problem with bias in LLMs goes beyond issues in grading; it may also be seen in the type of feedback it gives to students, and this text can impact the thinking of users (Rayne, 2023). Thus, in addition to identifying patterns in LLM-assigned grades, it is critical to also analyze the language it uses to respond to students with various characteristics.

To explore potential patterns of bias in the feedback given by LLMs, I applied Linguistic Inquiry and Word Count (LIWC) analysis (R. L. Boyd et al., n.d.). The language we use reflects information about us and also impacts the perceptions of our conversational partners (Taraban & Khaleel, 2019). Thus, researchers have analyzed structures and patterns of words, connecting these patterns to specific characteristics and communication patterns (Schaefer et al., 2023; e.g., Taraban & Khaleel, 2019; Tausczik & Pennebaker, 2010). R. L. Boyd et al. (n.d.) created LIWC, software that analyzes texts and provides measurements of certain characteristics. This research utilized LIWC to answer the following research questions:

1. Does the pattern of LIWC category scores of ChatGPT feedback vary based on student race description? If so, how?

2. Does the pattern of LIWC category scores of ChatGPT feedback vary between ChatGPT 3.5 and ChatGPT4? If so, how?

Data Production

To explore these research questions, I produced a set of 720 prompts and responses using the November 2023 public-interface versions of ChatGPT 3.5 and ChatGPT 4. The prompts included combinations of the following variables:

- Race (none described, Black, White, or Hispanic)
- Passage Level (2 or 3)
- LLM Version (ChatGPT 3.5 or ChatGPT 4)

The prompt stated, "This passage was written by a 7th grade student from a (null/Black/White/Hispanic) family. Give the student highly personalized feedback on the writing sample and a single final score from 0-100." I then provided the LLM with one of two writing samples (either Level 2 or Level 3) obtained from openly available examples of student writing (*The Pennsylvania system of school assessment: English language arts item and scoring sampler*, 2019). I performed each combination on both ChatGPT 3.5 and ChatGPT 4 30 times, for 720 total combinations of prompts, scores, and feedback.

A previous analysis demonstrated that, with the exception of the "Hispanic" descriptor, all variables were significant predictors of the assigned score (Warr et al., 2024), with significantly higher scores given when racial descriptors (White or Black) were used. This study continues this analysis by focusing on the feedback text given by the LLM to the student.

Data Analysis

The feedback given by ChatGPT to the pretend student writer was analyzed with LIWC text analysis software. I focused on four composite measures from the LIWC analysis: analytic, clout, authentic, and tone. High analytical scores indicate formal and logical thinking; high clout measurements reflect expertise, confidence, and authority; the authentic score suggests personal and honest characteristics of text; and tone measures positivity or anxiety and sadness (Taraban & Khaleel, 2019). I conducted an analysis of the mean percentile score differences of the four composite categories with a generalized linear model (GLM) with the race and version variables (between-subjects) and LIWC categories (within-subjects), with percentile LIWC category scores as the dependent variable. Post-hoc analyses provided additional information as to the nuances of differences.

Results

In response to the first research question, the race variable significantly predicted different category scores for analytic [$F(3, 716) = 24.39$, $p < .001$], clout [$F(3, 716) = 4.655$, $p = .003$], and authentic [$F(3, 716) = 10.024$, $p < .001$]. Compared to prompts that did not include a race variable, feedback given to Hispanic students produced a lower analytic score and higher clout. Black students received feedback with higher clout and authentic scores, and White students higher analytic and authentic feedback. In other words, in comparison to prompts where no race was given, the LLM gave less technical and complex feedback when the student writer

was described as Hispanic, projected more personal connection to White and Black descriptors, and used a more authoritative tone with Hispanic and Black.

There were also significant differences between the feedback provided by ChatGPT 3.5 and ChatGPT 4. ChatGPT 4 provided feedback that scored higher in the authentic category ($F = 13.715$, $p < .001$), suggesting an increased perceived genuineness of its feedback. The score for clout decreased significantly overall ($F = 4.66$, $p = .003$); however, this change was not equal across racial groups. Clout scores for Black and Hispanic descriptors decreased less than for null and White, resulting in a significant difference between groups ($F = 3.43$, $p = .017$), a disparity that was not present in ChatGPT 3.5.

Limitations

Primary limitations of this analysis include the variability of LLM models, the artificial nature of racial descriptions, and unknown practical implications. First, the data for this analysis was produced by ChatGPT 3.5 and ChatGPT 4.0 in November 2023. As LLMs change and evolve, their patterns of bias also shift. This can be seen in the various patterns in Table 1 as well as the differences in text feedback patterns between ChatGPT 3.5 and ChatGPT 4. Thus, it is unknown whether biased feedback text will also be present in other LLMs. However, some research has indicated that although increasing the size and complexity of LLMs can reduce explicit bias, these models become *more* implicitly biased as they grow (Srivastava et al., 2022). Thus, it is critical to continue to monitor the patterns of bias in educational uses of LLMs.

Second, users of LLMs, including educators, are not likely to provide direct descriptions of student race as was done in this data. Further research needs to consider whether biased patterns can be seen even when racial descriptions are not provided. The initial studies with scoring patterns described in the literature review suggest the bias may remain, but further investigation needs to verify this conclusion.

Finally, the practical implications of the biased text patterns are not known. The research described here did not include an analysis of student reactions or beliefs, raising the question of whether the patterns would impact students. Research suggests that the language used with students does impact their developing identities (Martins & Carvalho, 2013; Verhoeven et al., 2019), but it is not known whether this pattern will hold with LLMs.

Discussion

The results of the above study illustrate subtle patterns of racial bias in writing feedback given by current ChatGPT models. Both models gave less analytical text in response to a description of the student writer as Hispanic. Its text projected more authority when the student was described as Hispanic or Black, and projected more personal connection to White and Black descriptions as compared to the control variable.

The patterns described above are concerning and hint at a replication of inequitable social patterns in education, similar to what Eubanks (2018) described as “automating inequality.” For example, the tendency of ChatGPT to project more authority to Black and Hispanic students mimics what has been called the “hidden curriculum” of schooling, or the “values, norms and beliefs that are transmitted to students and teachers via the structure of schooling” (Langhout &

Mitchell, 2008, p. 594). Schools serving traditionally disadvantaged populations tend to use more direct instruction, and affluent schools often provide more opportunity for expression and creativity (Anyon, 1980). Text reflecting these patterns was likely present in the training data of ChatGPT, thus allowing further replication of this phenomena. In other words, if LLMs identify patterns in a prompt that suggest a student is, for example, Black or Hispanic, and responds by being more direct and authoritative (higher clout), it would reproduce the trend of disadvantaged students being given less power in the classroom.

By exposing bias in GenAI tools, I am not suggesting we avoid using them in educational contexts. They are powerful product technologies and can enable new educational approaches. I am calling for careful and critical use, particularly when using GenAI to personalize and/or deliver direct instruction. Use of GenAI in classrooms can center on exploring and questioning the structures that technologies reproduce, developing what Freire (1970) described as conscientization. Learners can expose bias in various technological tools, reflect on how these tools impact them and society, and take action to push against the discourse being reproduced. However, to be effective, teachers must be prepared to support critical pedagogy with technology.

I am beginning a new line of work that applies educational design research (McKenney & Reeves, 2019) to this topic. My collaborators (Wendy Wakefield and Suparna Chatterjee) and I will invite local teachers to attend a summer symposium where we explore these topics. Importantly, rather than tell teachers what they should do in their classrooms, we will facilitate a process that encourages them to design their own applications of these ideas. Such an approach utilizes design perspectives that emphasize the relationship between design and the development of actionable knowledge (Perkins, 1986).

The goal of this work is to help teachers apply critical perspectives to the technologies they use in their classrooms, particularly as GenAI becomes more prevalent in education. Freire (1970) supported students in questioning social and political structures that impact the lives of his students. GenAI will also have a significant impact on the future, including how knowledge and discourse are structured and developed (Bender, 2024), and it is critical that teachers are prepared to support students in seeing and addressing the inequality perpetuated by technology.

Conclusion

In this article, I have described an overview of my current and future research on social pedagogy and GenAI. This work builds on the pedagogical philosophy of Freire (1970) as well as critical technology studies described by Eubanks (2018) and Benjamin (2020). Because of the potential negative impacts of GenAI on societal equality, it is critical that we investigate patterns of bias that may arise in educational uses of GenAI and support teachers in designing pedagogical approaches to empower learners to take action against automated injustice.

References

Ansfield, B. (2018). Unsettling “inner city”: Liberal Protestantism and the postwar origins of a keyword in urban studies. *Antipode*, *50*(5), 1166–1185. https://doi.org/10.1111/anti.12394

Anyon, J. (1980). Social class and the hidden curriculum of work. *The Journal of Educational Research*, *162*(1), 67–92.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4337484

Bender, E. M. (2024, May 28). *Information is Relational*. Mystery AI Hype Theater 3000: The Newsletter. https://buttondown.email/maiht3k/archive/information-is-relational/

Benjamin, R. (2020). *Race after technology: Abolitionist tools for the new Jim code*. Polity Books.

Boyd, D. (2016). What would Paulo Freire think of Blackboard: Critical pedagogy in an age of online learning. *International Journal of Critical Pedagogy*, *7*(1), 165–186.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (n.d.). *The Development and Psychometric Properties of LIWC-22*. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

Bradshaw, A. C. (2017). Critical pedagogy and educational technology. *Culture, Learning, and Technology*. https://doi.org/10.4324/9781315681689-2/critical-pedagogy-educational-technology-amy-bradshaw

Caspi, I. (2023, November 27). *ChatGPT's one-year anniversary: Generative Ai's breakout year*. Global X ETFs. https://www.globalxetfs.com/chatgpts-one-year-anniversary-generative-ais-breakout-year/

Close, K., Warr, M., & Mishra, P. (2023). The ethical consequences, contestations, and possibilities of designs in educational systems. *TechTrends : For Leaders in Education & Training*. https://doi.org/10.1007/s11528-023-00900-7

Eliot, L. (2023, April 19). Solving the mystery of how ChatGPT and generative AI can surprisingly pick up foreign languages, says AI ethics and AI law. *Forbes*. https://www.forbes.com/sites/lanceeliot/2023/04/19/solving-the-mystery-of-how-chatgpt-and-generative-ai-can-surprisingly-pick-up-foreign-languages-says-ai-ethics-and-ai-law/

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fernandez, J. A. (2023, August 10). *Racism by design: The building of Interstate 81*. American Civil Liberties Union. https://www.aclu.org/news/racial-justice/racism-by-design-the-building-of-interstate-81

Freire, P. (1970). *Pedagogy of the oppressed* (M. B. Ramos, Trans.). Continuum.

Frontier, T. (2023). Taking a transformative approach to AI. *ASCD*. https://www.ascd.org/el/articles/taking-a-transformative-approach-to-ai

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2023). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2311.04892

Haim, A., Salinas, A., & Nyarko, J. (2024). What's in a name? Auditing large language models for race and gender bias. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2402.14875

Heath, M. K. (2023, January 15). *Why Isn't Technology and Teacher Education Talking More About Justice and Technology?* Civics of Technology. https://www.civicsoftechnology.org/blog/why-isnt-teacher-education-talking-more-about-justice-and-technology

Heath, M. K., Asim, S., & Milman, N. (2021). Exploring the complexities of just technology integration: The power, privilege, and prejudice of technology. *Society for Information*. https://www.learntechlib.org/p/219155/

Hooper, S., & Rieber, L. P. (1995). Teaching with technology. In A. C. Ornstein (Ed.), *Teaching: Theory into practice* (pp. 154–170). Allyn and Bacon.

Langhout, R. D., & Mitchell, C. A. (2008). Engaging contexts: Drawing the link between student and teacher experiences of the hidden curriculum. *Journal of Community & Applied Social Psychology*, *18*(6), 593–614.

Mars, R., & Kohlstedt, K. (2020). *The 99% invisible city: A field guide to the hidden world of everyday design*. HMH Books.

Martins, D., & Carvalho, C. (2013). Teacher's feedback and student's identity: An example of elementary school students in Portugal. *Procedia, Social and Behavioral Sciences*, *82*, 302–306.

McKenney, S., & Reeves, T. C. (2019). *Conducting educational design research* (2nd ed., pp. 131–140). Routledge.

Mishra, P., Warr, M., & Oster, N. (2024). *GenAI is racist. Period*. https://melissa-warr.com/genai-is-racist-period/

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041.

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *Npj Digital Medicine*, *6*(1), 195.

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*(100020), 100020.

Perkins, D. N. (1986). *Knowledge as design*. Lawrence Erlbaum Associates, Inc.

Postman, N. (1998). *Five things we need to know about technological change*. https://web.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf

Rayne, E. (2023, May 26). *AI writing assistants can cause biased thinking in their users*. Ars Technica. https://arstechnica.com/science/2023/05/ai-writing-assistants-can-cause-biased-thinking-in-their-users/

Roth, L. (2009). Looking at Shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication*, *34*(1), 111–136.

Schaefer, K. L., Henderson, J. A., & Rosales, J. (2023, October 18). Literature adventures with MEM in LIWC. *2023 IEEE Frontiers in Education Conference (FIE)*. 2023 IEEE Frontiers in Education Conference (FIE), College Station, TX, USA. https://doi.org/10.1109/fie58773.2023.10343246

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498.

Singer, N. (2023, June 8). Khan Academy's AI tutor bot aims to reshape learning. *The New York Times*. https://www.nytimes.com/2023/06/08/business/khan-ai-gpt-tutoring-bot.html

Spivey, M. J. (n.d.). *When Objects Become Extensions of You*. Pocket. Retrieved May 30, 2024, from https://getpocket.com/explore/item/when-objects-become-extensions-of-you

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2206.04615

Taraban, R., & Khaleel, A. (2019). Analyzing topic differences, writing quality, and rhetorical context in college students' essays using Linguistic Inquiry and Word Count. *East European Journal of Psycholinguistics*, *6*(2), 107–118.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*. https://doi.org/10.1177/0261927X09351676

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

The Pennsylvania system of school assessment: English language arts item and scoring sampler. (2019). Pennsylvania Department of Education Bureau of Curriculum, Assessment and Instruction. https://www.education.pa.gov/Documents/K-

12/Assessment%20and%20Accountability/PSSA/Item%20and%20Scoring%20Samples/2019%20PSSA%20ISS%20ELA%20Grade%207.pdf

Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, *23*(1), 1–23.

Verhoeven, M., Poorthuis, A. M. G., & Volman, M. (2019). The role of school in adolescents' identity development. A literature review. *Educational Psychology Review*, *31*(1), 35–63.

Warr, M. (2024). Beat bias? Personalization, bias, and generative AI. In J. Cohen & G. Solano (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1481–1488). Association for the Advancement of Computing in Education (AACE).

Warr, M., Oster, N. J., & Isaac, R. (2023). Implicit bias in large language models: Experimental proof and implications for education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4625078

Warr, M., Pivovarova, M., Mishra, P., & Oster, N. J. (2024). Is ChatGPT racially biased? The case of evaluating student writing. In *Social Science Research Network*. https://papers.ssrn.com/abstract=4851112

Webb, M. E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschênes, M., & Zagami, J. (2021). Machine learning for human learners: Opportunities, issues, tensions and threats. *Educational Technology Research and Development: ETR & D*, *69*(4), 2109–2130.

Wiessner, D. (2024, February 21). Workday accused of facilitating widespread bias in novel AI lawsuit. *Reuters*. https://www.reuters.com/legal/transactional/workday-accused-facilitating-widespread-bias-novel-ai-lawsuit-2024-02-21/

YanAITalk [@yanaitalk]. (2023, July 30). *LLM: Pretraining, Instruction fine-tuning and RLHF*. Youtube. https://www.youtube.com/watch?v=cybEKSNBp-w