**Challenging AI Bias through Critical Pedagogy**

Melissa Warr

New Mexico State University, New Mexico, USA

warr@nmsu.edu

**Abstract:** In this paper, I explore progress in research on the intersection of Generative Artificial Intelligence (GenAI) and critical pedagogy, advocating for an equity-oriented approach to AI in education. I describe four guidelines for critical AI use: active engagement, continual reflection, collaborative exploration, and creative discovery. Then I describe two contexts in which I am studying how these guidelines impact AI use: a study with pre-service teachers, and a second study with local in-service teachers that includes teachers designing curriculum for other teachers.

Keywords: Large language models (LLMs), critical pedagogy, in-service teacher education, pre-service teacher education

Brazilian educator Paulo Freire empowered learners to develop critical consciousness—to recognize and question societal inequities and understand the evolving nature of social structures [(Freire, 1972)](https://paperpile.com/c/bW17kv/GZ7R1). Freire believed that education should foster awareness of inequities and injustice, enabling individuals to challenge and transform their world. This concept has gained renewed significance as Generative Artificial Intelligence (GenAI) increasingly influences educational environments. By affecting how we access, interpret, and validate knowledge [(Bender, 2024)](https://paperpile.com/c/bW17kv/Qw23I), GenAI impacts whose perspectives are amplified or marginalized.

Throughout history, technological shifts have redefined knowledge systems and societal organization, reinforcing dominant social values [(Postman, 1998; Spivey, n.d.)](https://paperpile.com/c/bW17kv/CR9IQ%2BXLbyR) This dynamic is evident in the biases embedded within GenAI. GenAI tends to present information as fixed and objective "truth," disregarding the nuanced, fluid, and evolving social realities that Freire (1972) emphasized. Recent research, including that described in my chapter in RHTTE, has highlighted how large language models (LLMs) adjust grading and feedback in response to learner descriptions [(Warr, 2024)](https://paperpile.com/c/bW17kv/LhLLN). For example, if a learner is labeled as Hispanic or Black, LLMs provide feedback that is more direct and authoritative than a learner labeled as White. Addressing this inequity calls for critical approaches for AI in education.

I use the term “critical approaches” or “critical uses of AI” to describe using AI from an equity-oriented perspective. Rather than seeing its responses as a fixed truth, critical use perceives the generalizations and biases embedded in its responses, applying meta-cognition to monitor how its language impacts the user. This follows the Freirean principles of *conscientization* (critical consciousness, or seeing the inequities and contradictions in the world), praxis (reflection leading to transformation) and dialogue (ongoing reflection and discussion of inequitable experiences) [(Boyd, 2016; Bradshaw, 2017; Freire, 1972)](https://paperpile.com/c/bW17kv/Fl7Vh%2BGZ7R1%2BmrzNQ).

**Guidelines for Critical Use of GenAI**

Through my experiences teaching and speaking with and about GenAI, I have developed basic guidelines for educational uses of GenAI that may support critical use. In the spirit of educational design research [(McKenney & Reeves, 2020)](https://paperpile.com/c/bW17kv/bYoXV), the guidelines form an initial design or framework that can be further developed and refined through formal research.Table 1 outlines these four guidelines: active engagement, continual reflection, collaborative exploration, and creative discovery.

Table 1. Guidelines for Critical AI Use

|  |  |  |  |
| --- | --- | --- | --- |
| **Guideline** | **Description** | **Reasoning** | **Examples for Teacher Education** |
| Active Engagement | Mentally engage and practice skepticism while using AI. | It is easy to fall into passive use of AI, leading to making assumptions of its truth and accuracy as well as limiting learning. Not only does this include accepting factual inaccuracies, but also a lack of reflection on the type of reality the AI is presenting, which often includes embedded biases and including some perspectives more than others. | When designing GenAI activities, include elements that require learners to fully engage in the content, not simply scan text. For example, a chatbot can be instructed to wait until a learner has fully responded to its question before proceeding.Additionally, learners can be asked to not just fact-check information, but identify assumptions the LLM made and where it “filled in the blanks,” adding content that was not requested. Activities with creating images are particularly effective for observing these differences (see below 1000 words activity described below). |
| Continual Reflection | Reflect during and after AI interactions; practice metacognition | It's crucial to be mindful of our thoughts and reactions because it's easy to fall into the trap of believing what we are told by LLMs. Some GenAI tools, like LLMs, are designed to appear human-like, but they have odd quirks and idiosyncrasies. Our brains naturally tend to believe what we're told; it takes a conscious effort to avoid this and requires us to reflect on our own thoughts to catch assumptions. | Ask learners to complete AI “think-alouds”: as they use a GenAI tool, record their screen and verbally state their thoughts. Then asks learners to watch their interaction and notice when they were making assumptions as well as how the AI impacted their thinking.Learners can also write or record brief reflections after they complete an activity. |
| Collaborative Exploration | Use GenAI tools with others, having conversations in the process | GenAI tools can seem human, but they are not. By exploring AI tools with others, learns become familiar with the differences between AI responses and human responses. Conversing with other humans can also deepen our reflection and metacognition and encourage active engagement. | When completing an AI activity, first complete the task with the whole class, discussing thoughts and observations throughout. Model metacognition and reflection. Then ask learners to do the task in small groups or pairs, and listen to their conversations to encourage critical thinking. |
| Creative Discovery | Experiment and play | GenAI tools are unique and we do not know what all they can do. To come to understand the nature of these tools, try using them for different tasks, find the things they do well and the things they do poorly, and just be creative! | Conduct “quick fires”: time-limited challenges that are impossible to do, but perhaps more possible if using GenAI. For example, ask learners to create a three culturally responsive storybooks at multiple levels. Always reflect on what worked and what didn’t and what biases can be seen. |

**Current Research Approaches for Critical AI Use**

I am currently applying these four guidelines in two core research projects that explore critical pedagogy and AI, one with pre-service teachers and one with in-service teachers.

**Experimentation with Pre-Service Teachers**

As a small step towards exploring how to develop critical uses of GenAI, I have been experimenting with how to help pre-service teachers develop these skills through various activities and prompts. Central to this work has been leading collaborative activities and reflective discussions that encourage critical thinking about AI.

For example, I have created a bank of LLM prompts that we experiment with in class. The prompts create a chatbot that interacts with the learner around some topic. I encourage critical perspectives by following this general activity structure:

1. Try the prompt as a whole class, discussing what we notice about the LLM responses and why it responds as it does
2. Continue experimenting, working in small groups or pairs and discussing what they think and notice
3. Share out their observations with the full class
4. Write or record a brief reflection about their experience, including what they were thinking during the activity

These activities are not only for AI practice, they are also integrated into the course content. For example, at the beginning of the semester we talked about unit and lesson planning, discussing how we can create authentic, technology-enabled assessments to assess learning standards. I created an LLM prompt that helped learners connect the ideas of curriculum standards with authentic assessments. It asked learners for a content area and subject they wanted to explore, then gave them a variety of ideas about authentic assessments that might evaluate those standards. The activity not only exposed learners to AI, but also provided scaffolding that helped them shift their mindset from assessments as tests to more authentic creations.

Although this activity did not show obvious or major biases, it still provided opportunities to discuss how the LLM’s training data impacted it’s responses. For example, we discussed how the types of activities it suggested were often US or Western-centric, such as suggestions to explore societal structures through democracy role plays. The procedure of doing it as a whole class, then in groups/pairs supported conversation about what we were thinking and doing, building the cognitive habits that may lead to critical AI use.

Other prompts explored connecting standards and tech tools and understanding how LLMs predict next tokens in often biased ways. A list of prompts (many still in development) can be found [here](https://docs.google.com/document/d/1qouqVxCvG-zeWMYNf8Y12Jb23yayPD5T_DtiRXApFbQ/edit?usp=sharing).

I have also observed that one of the best ways to understand the properties of GenAI is through visual activities. For example, we explored how AI “fills in the blanks” with information through the activity “1000 AI Words,” originally designed by Nicole Oster, a doctoral student at Arizona State University:

1. Select a picture of themselves or something else about them—this could be one they took or just an image from the internet. It’s also fun to follow this process with a quick picture of the class.
2. Ask an LLM to write a 1000 word description of the image.
3. Copy the 1000 word description into a new chat.
4. Ask the LLM to create an image based on the 1000 words.
5. This description was copied into a new chat with instructions to create a new image based on the words.

The results of the experiment were eye-opening: we became models and stock photos! While AI can accurately depict many things, it also reveals the nature of its training data and its patterns of generalization. It's clear that the AI's training data must have contained a significant amount of models and stock photos, as that's primarily what it generates. We also observed the AI's tendency to add elements to a scene and make it more generic. For instance, in one case, it added an American flag to a classroom, and in another, it provided every student with a laptop. Additionally, the AI tends to "whitewash" images, presenting people as predominantly white. Comparing the images visually demonstrates how AI relies on its training data rather than the real world, and these differences are more easily perceived visually than through textual descriptions.

These small experimental activities have provided ideas for what might be done in a K-12 classroom to promote critical AI use. However, teachers in those classrooms are most suited to create activities for their learners. I am also beginning a major research study that collaborates with local teachers in promoting critical perspectives on AI.

**Educational Design Research Study with Local In-Service Teachers**

I am beginning a research project that applies Educational Design Research as described by McKenney and Reeves (2020). This project has received three sources of funding: a racial equity grant from the Spencer Foundation, an internal New Mexico State University (NMSU) Emerging Scholars Grant, and a generous donation by Richard Leza, an NMSU alumni. The project has three phases:

1. **Leza Scholars**: a collaboration with 6 local teachers to:
	* Experiment with curriculum and activities that support critical AI use in various content areas
	* Design a summer symposium for a larger group of teachers
	* Create a website/blog to share our ideas and learning designs
2. **Summer Symposium**: Two-week summer symposium to be held in June 2025. This symposium will invite 10-20 teachers to the New Mexico State University campus to learn about critical approaches to AI and design activities for their classrooms. The Leza Scholars will lead the symposium and support their fellow teachers
3. **Follow-up**: Ongoing meetings with participants in Fall 2025 to see how our work together impacts their classroom use of AI

The hope is that through these collaborations, we are able to both serve teachers and learners in our area as well as develop theory and practices that can support critical uses of AI.

**Conclusion**

Generative AI is an exciting field of study—there are many new areas to explore. The tools are constantly changing, sometimes improving and sometimes raising cause for concern. Given how rapidly these tools are entering classrooms, it is critical that we evaluate what impact it may have on educational equity as well as design methods to ameliorate potential negative impacts. My research focuses on combining critical pedagogy with GenAI to bring attention to these important issues.

**References**

[Bender, E. M. (2024, May 28). *Information is Relational*. Mystery AI Hype Theater 3000: The Newsletter.](http://paperpile.com/b/bW17kv/Qw23I) <https://buttondown.email/maiht3k/archive/information-is-relational/>

[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.](http://paperpile.com/b/bW17kv/Fl7Vh)

[Bradshaw, A. C. (2017). Critical pedagogy and educational technology. *Culture, Learning, and Technology*. https://doi.org/](http://paperpile.com/b/bW17kv/mrzNQ)[10.4324/9781315681689-2/critical-pedagogy-educational-technology-amy-bradshaw](http://dx.doi.org/10.4324/9781315681689-2/critical-pedagogy-educational-technology-amy-bradshaw)

[Freire, P. (1972). *Pedagogy of the oppressed* (M. B. Ramos, Trans.).](http://paperpile.com/b/bW17kv/GZ7R1)

[McKenney, S., & Reeves, T. C. (2020). Educational design research: Portraying, conducting, and enhancing productive scholarship. *Medical Education*. https://doi.org/](http://paperpile.com/b/bW17kv/bYoXV)[10.1111/medu.14280](http://dx.doi.org/10.1111/medu.14280)

[Postman, N. (1998). *Five things we need to know about technological change*.](http://paperpile.com/b/bW17kv/XLbyR) <https://web.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf>

[Spivey, M. J. (n.d.). *When Objects Become Extensions of You*. Pocket. Retrieved May 30, 2024, from](http://paperpile.com/b/bW17kv/CR9IQ) <https://getpocket.com/explore/item/when-objects-become-extensions-of-you>

[Warr, M. (2024). Blending generative AI, critical pedagogy, and teacher education to expose and challenge automated inequality. In R. Blankenship & T. Cherner (Eds.), *Research Highlights in Technology and Teacher Education*. AACE.](http://paperpile.com/b/bW17kv/LhLLN)