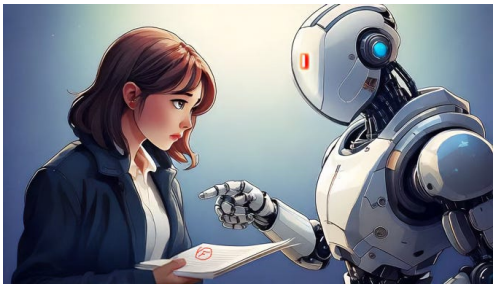

Accusatory AI: How a Widespread Misuse of AI Technology Is Harming Students

What should be done when an AI accuses a student of misconduct by using AI?

Publication note
Originally published in Towards Data Science. Author's copy on Substack.

Anti-cheating tools that detect material generated by AI systems are widely being used by educators to detect and punish cheating on both written and coding assignments. However, these AI detection systems don't appear to work very well and they should not be used to punish students. Even the best system will have some non-zero false positive rate, which results in real human students getting F's when they did in fact do their own work themselves. AI detectors are widely used, and falsely accused students span a range from grade school to grad school.

In these cases of false accusation, the harmful injustice is probably not the fault of the company providing the tool. If you look in their documentation then you will typically find something like:

> *"The nature of AI-generated content is changing constantly. As such, these results should not be used to punish students. … There always exist edge cases with both instances where AI is classified as human, and human is classified as AI."*
> — *Quoted from GPTZero's FAQ.*

In other words, the people developing these services know that they are imperfect. Responsible companies, like the one quoted above, explicitly acknowledge this and clearly state that their detection tools should not be used to punish but instead to see when it might make sense to connect with a student in a constructive way. Simply failing an assignment because the detector raised a flag is negligent laziness on the part of the grader.

If you're facing cheating allegations involving AI-powered tools, or making such allegations, then consider the following key questions:

- What detection tool was used and what specifically does the tool purport to do? If the answer is something like the text quoted above that clearly states the results are not intended for punishing students, then the grader is explicitly misusing the tool.
- In your specific case, is the burden of proof on the grader assigning the punishment? If so, then they should be able to provide some evidence supporting the claim that the tool works. Anyone can make a website that just uses an LLM to evaluate the input in a superficial way, but if it's going to be used as evidence against students

then there needs to be a formal assessment of the tool to show that it works reliably. Moreover this assessment needs to be scientifically valid and conducted by a disinterested third party.

- In your specific case, are students entitled to examine the evidence and methodology that was used to accuse them? If so then the accusation may be invalid because AI detection software typically does not allow for the required transparency.
- Is the student or a parent someone with English as a second language? If yes, then there may be a discrimination aspect to the case. People with English as second language often directly translate idioms or other common phrases and expressions from their first language. The resulting text ends up with unusual phrases that are known to falsely trigger these detectors.
- Is the student a member of a minority group that makes use of their own idioms or English dialect? As with second-language speakers, these less common phrases can falsely trigger AI detectors.
- Is the accused student neurodiverse? If yes, then this is another possible discrimination aspect to the case. People with autism, for example, may use expressions that make perfect sense to them, but that others find odd. There is nothing wrong with these expressions, but they are unusual and AI detectors can be triggered by them.
- Is the accused work very short? The key idea behind AI detectors is that they look for unusual combinations of words and/or code instructions that are seldom used by humans yet often used by generative AI. In a lengthly work, there may be many such combinations found so that the statistical likelihood of a human coincidentally using all of those combinations could be small. However, the shorter the work, the higher the chance of coincidental use.
- What evidence is there that the student did the work? If the assignment in question is more than a couple paragraphs or a few lines of code then it is likely that there is a history showing the gradual development of the work. Google Docs, Google Drive, and iCloud Pages all keep histories of changes. Most computers also keep version histories as part of their backup systems, for example Apple's Time Machine. Maybe the student emailed various drafts to a partner, parent, or even the teacher and those emails form a record incremental work. If the student is using GitHub for code then there is a clear history of commits. A clear history of incremental development shows how the student did the work over time.

To be clear, I think that these AI detection tools have a place in education, but as the responsible websites themselves clearly state, that role is not to catch cheaters and punish students. In fact, many of these websites offer guidance on how to constructively address suspected cheating. These AI detectors are tools and like any powerful tool they can be great if used properly and very harmful if used improperly.

If you or your child has been unfairly accused of using AI to write for them and then punished, then I suggest that you show the teacher/professor this article and the ones that I've linked to. If the accuser will not relent then I suggest that you contact a lawyer about the possibility of bringing a lawsuit against the teacher and institution/school district.

Despite this recommendation to consult an attorney, I am not anti-educator and think that good teachers should not be targeted by lawsuits over grades. However, teachers that misuse tools in ways that harm their students are not good teachers. Of course a well-intentioned educator might misuse the tool because they did not realize its limitations, but then reevaluate when given new information.

> *"it is better 100 guilty Persons should escape than that one innocent Person should suffer" — Benjamin Franklin, 1785*

As a professor myself, and I've also grappled with cheating in my classes. There's no easy solution, and using AI detectors to fail students is not only ineffective but also irresponsible. We're educators, not police or prosecutors. Our role should be supporting our students, not capriciously punishing them. That includes even the cheaters, though they might perceive otherwise. Cheating is not a personal affront to the educator or an attack on the other students. At the end of the course, the only person truly harmed by cheating is the cheater themself who wasted their time and money without gaining any real knowledge or experience. (Grading on a curve, or in some other way that pits students against each other, is bad for a number of reasons and, in my opinion, should be avoided.)

Finally, AI systems are here to stay and like calculators and computers they will radically change how people work in the near future. Education needs to evolve and teach students how to use AI responsibly and effectively. I wrote the first draft of this myself, but then I asked an LLM to read it, give me feedback, and make suggestions. I could probably have gotten a comparable result without the LLM, but then I would likely have asked a friend to read it and make suggestions. That would have taken much longer. This process of working with an LLM is not unique to me, rather it is widely used by my colleagues. Perhaps, instead of hunting down AI use, we should be teaching it to our students. Certainly, students still need to learn fundamentals, but they also need to learn how to use these powerful tools. If they don't, then their AI-using colleagues will have a huge advantage over them.

About Me: James F. O'Brien is a Professor of Computer Science at the University of California, Berkeley. His research interests include computer graphics and animation, simulation of physical systems, human perception, rendering, artificial intelligence, computer vision, virtual reality, digital privacy and security, and image and video forensics.

If you would like to read my other articles, consider subscribing. You can also find me on LinkedIn, Instagram, and at UC Berkeley.