Essay Draft

Lucas Vas

AI has made leaps and bounds over the past several years. It's starting to become a part of life, something that we all should understand how to use, at least on a very basic level. Its capabilities start at something as simple as summarizing a Google search, to writing code for a program or webpage that looks good and functions (relatively) well. It can "understand" information that's fed to it from every field - the medical field, the math field, sports, judicial branch - if there's text to read, then it can interpret it in some way. AI is quite the tool, by far one of the most industry-shaking technological advancements in the 21st century, but it still has its issues. One of the largest issues with AI, in its current state, is that no one really understands how it works. Not one person can trace the decision-making process back to something we can understand as humans, however we still allow the AI to influence our own creative process and understanding of the world. I believe that the lack of explainability within AI models, with a focus on large language models that have been used to shape our decisions, raises serious ethical concerns due to the lack of fairness, accountability, and transparency, which leads to a lack of trust and holds back fully responsible deployment of AI into our everyday lives.

1 Transparency

1.1 What is transparency?

To start, what is transparency? The word transparency was derived from the Latin word *transparentia*, which translates to "shining through." It literally means that you can see through something, like a window or a plastic covering. In business, this typically has to do with the policies that are being created or have been created already. When a business creates a policy, everyone that interacts with said company must follow that policy, otherwise they will suffer some sort of consequece. Whether this is some severe consequence or something small, it will affect people and how they interact. To apply this principle of transparency to policy-making, companies will generally hold meetings that are accessible to the public in some way. In recent times, this could look like an online meeting in the form of a Zoom call, it could be a recorded in-person meeting, or any other type of record keeping. This creates the ability for people that aren't directly involved in the meetings, such as a customer or a lower-ranked employee, to understand the process that led up to the creation of a policy. Note that this has nothing to do with the people's ability to debate the decision, assuming they find it unjust - this is purely so that people understand why decisions are being made and how they may affect them.

1.2 How does this apply to AI?

When we start applying this principle to AI, it looks quite similar. We still want to be able to understand the creation of the decisions that affect us. We still want to be able to have some inkling of the process that the AI uses to find its conclusions. This is where the issue with the AI comes into play - there's no functional way of asking the AI how it came to its conclusion. One can think of the AI as a pathological liar - it will lie with so much confidence that you wouldn't know it's wrong until you attempt to verify the information. There are plenty of cases of AI givng the wrong answers to questions that people have asked it. Asking the AI about specific people and places will sometimes result in wildly incorrect answers. The issue is that we don't always know where these answers are coming from. The second part of this issue is that you can't simply ask the AI how it came up with the answers - this goes back to the pathological lying example that was given earlier. The AI is designed, at its most basic levels, to generate words that look good next to each other and in context to the previous information generated, but it doesn't necessarily "understand" that information in the way that humans do. Therefore, you can't necessarily trust any of the information that it puts out to you and there's still quite a bit of verification that needs to go into the repsonse that you recieve. This still applies when you ask the AI how it created any response. It will print out a response quickly and easily, and usually it will look like a good explanation. The issue with this is that since you can't trust the AI's responses in the first place, you also cannot trust it to explain itself because that could be just as inconsistent with real information. This is a classic "boy who called wolf" situation.

Another good example of this is discussed in the paper "GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems." In this paper, the authors discuss the use of GPT-4 to answer questions that are considered to be what computer scientists call NP-complete problems. These problems are very difficult for computers to solve, and, given a large enough dataset, have been proven to be impossible to solve in a reasonable amount of time. The authors of this paper used GPT-4 to answer these questions, and found that it was able to give them answers quickly, but not quite accurately. As they said in the article:

> "At some point in the backprompts of 40 instances, the generating model returned an optimal coloring. In none of those instances did the verifying GPT realize this. In 39 cases, it hallucinated pairs of vertices that it claimed were adjacent and same-colored." [2]

As they explained, they used a separate model to supply the AI with correct answers to the questions that they were asking, which in this case was the “graph coloring” problem. GPT, which was being used to verify the answers by generating its own answers, was unable to verify the answers that it was given, hallucinating answers that were completely incorrect. This is a perfect example of the AI being unable to explain itself, and also being unable to verify its own answers. This is a problem that is very difficult to solve, and is one of the main reasons that AI is not transparent.

Another part of transparency is its integration with privacy. As humans, we tend to value our privacy, to the point of buying into systems designed to protect that at any cost. This doesn’t change much when digital privacy enters the equation. Looking at AI through this lens shows us a couple possibly breaches of privacy. Large language models specifically, or LLM’s, have to be trained on huge datasets. These datasets can be anything in the public domain - news articles, books, open source codebases, etc. They can also be more private data - social media posts, texts, recorded speech, pictures. To the AI, you’re just a number associated with a series of data entries. The more data the AI is fed, the more data it can relate directly to you.

Why does this matter? With respect to transparency, privacy is one of the reasons we value transparency so much. Since transparency governs how much we know about what happens with decisions, it follows that it also governs how much we know about our own privacy. Having AI be completely opaque, the way it is right now, means that we don’t have any idea how much our personal information is being used. This is bad. For all we know, some giant corporations - Google, Microsoft, and Meta to name a few - use our information to train their AI models. The issue isn’t necessarily that we don’t know they use our information, let’s be honest, that’s no secret. The true issue stems from the fact that we don’t know what they’re doing with that information. They could be creating a massive model that simply uses the information to “talk” a little bit more like a human, or (put on your tinfoil hats) they could be creating a giant model that will run governments in a way that allows them to do whatever they want. The bottom line is that no one knows for sure what these corporations are doing with the data that they've taken.

2 Accountability

2.1 Definition of accountability

Accountability is just as important as anything else. We tend to combine accountability with responsibility, and then use the words interchangeably. I'd like to think that this is because that's exactly how closely related they are. In the case of ethics, the definition of the principle of accountability is almost identical to Merriam-Webster's definition of responsibility: "able to answer for one's conduct and obligations." In terms of AI, this means that the AI should be able to answer for anything that it creates or says. Accountability is a fairly important principle, especially easy to see in court and the medical field.

2.2 How does accountability relate to AI?

When we take a look at this in those fields, we see examples with a person making assertions to another person. This person should, in theory, be more knowledgable about the subject than the second person receiving the advice. For example, in the medical field, a medical practitioner, be that a surgeon, doctor, or nurse, should be able to make a diagnosis of the patient. Said patient should give their symptoms to the doctor, and then receieve their diagnosis. However, the patient does have the option to appeal to the doctor and say that their diagnosis was incorrect. This applies almost the same way to the judicial system. The judge will have all of the information laid out in front of them, and they will create a ruling based off of all that information. The defendent and/or prosecutor both still have the option to debate this ruling, to a certain extent. Obviously there's other limitations due to the fact that it's a legal matter. In both of these examples, the "more knowledgable" person (the judge and doctor) are making decisions that affect the "less knowledgable" people (defendent and patient), however the judges and doctors must still be accountable for the rulings or diagnoses that they've issued. This is where AI fails to meet expectations and requirements for responsibility. Take another example, this time using an AI driven car - something close to a Tesla, with its "hands-off driving" modes, but with even less user interaction. Assuming the AI controls this car totally and completely, we would assume that it is also fully responsible for its own actions while exerting this control. However, this is not so. If the car is involved in an accident of any sort, either involving another car or (God forbid) a pedestrian, the AI is in no way legally responsible for that accident. We should also remember that the accident can cause the death of a person, which in many court cases results in the survivor, who allegedly caused the accident, being charged with a crime such as vehicular manslaughter and negligent driving. In our hypothetical car accident, the person that was in the driver seat at the time may be the person charged with these crimes, simply because the AI cannot be held responsible for its own actions.

In her article, "AI, Explainability, and the Human Mind", Jocelyn Maclure discusses the idea of accountability and explainability in AI. She specifically references the healthcare industry, and how AI is being used to diagnose patients. She says this about the accountability of AI:

> "In healthcare, for instance, a diagnosis, prognosis, or treatment recommendation made by a deep learning algorithm should be confirmed by further medical testing or by a physician's clinical judgement." [1]

She also touches upon the same idea that I mentioned earlier with the judicial system:

> "In the judicial system, a judge should always be able to explain and justify a particular sentence or why bail or parole is granted or not. This is a concrete way to give substance to the vague mantra that humans ought to be 'kept in the loop.'" [1]

As we can see, Maclure agrees and expands upon the idea that AI should be accountable for its actions. She mentions earlier in her article that AI is not able to explain itself, and that although ”it looks as if the improved performance enabled by AI justifies relxaing the explainability requirement, it is actually not clear that it is so.” [1] We agree that AI is a powerful tool, but also that it is not quite ready to be used in certain fields.

As I mentioned while discussing transparency, AI cannot create an explanation as to how or why it reached any given conclusion. It constantly has no understanding of what it's saying or why it might be saying what it is. Therefore, there's no way that AI can be accountable for any of its decisions or results. The true problem with this isn't that it can't be explained right now, but more so that the AI is still being used in various places. The doctor that was giving you your diagnosis or the judge sentencing you to 25 years in prison are being influenced by technology that cannot be fully explained. You would no longer be able to appeal to the doctor and say that the diagnosis is wrong, because the doctor is just the AI. You would never be able to ask the judge why they issued their ruling because the AI couldn't explain itself either.

There are some great examples of this happening all over the internet. To name one, there have been studies done that have proven that AI can be extremely biased. Since the AI must be trained on some sort of data, it follows that its data could also be biased, driving the AI in a direction of bias. The problem here is that even when fed what we would consider non-biased data, the AI may still end up becoming biased towards a specific group of individuals, or, more specifically, the majority of black people. However, this isn't a discussion about the presence of racism in the media or general populace.

3 Fairness

3.1 What is fairness?

Fairness, as a concept, is quite hard to define. There's multiple different definitions that have completely different applications depending on what you're applying it to. There's also what I would consider to be "levels" of fairness. These "levels" are created as a direct result of the fact that, in some cases, one can never be "completely" fair. A couple of the examples that we've discussed in class are closely related to the distribution of resources, especially in a capitalistic society, where your gains are almost directly related to how much work you've put in. There's always been a logistical and ethical debate, especially in politics, centered around the distribution of wealth in a society that people argue is, by definition, unfair. On top of this, there's the implicit discrimination of black people in our society, creating an even larger divide. When we attempt to relate this to AI, there are a couple similarities. Just like the arguments involving people of color in our society, access to AI is incredibly discriminatory. Not everyone in the world is able to access AI, whether for financial reasons, racial reasons, etc. Not everyone has access to the technology prerequisites in order to understand how it works, or to be able to get to the website that hosts our interfaces into various models. Past this, you can enter into the realm of competition with the AI corporations - there's little to no way to compete with these gigantic companies. Since the power of an AI is directy linked to how much data it has access to and how much processing power is behind it, someone starting a new company has almost no way to compete with the corporations, simply due to the fact that they do not have access to the base requirements of the AI's processing power. This could be seen as a violation of fairness.

3.2 Example of fairness in AI

A direct example of the violation of this principle is in a hypothetical situation. An age old debate in the world of politics, even on the local level, is the debate of whether or not to create some sort of predictive policing system. This system would use AI to predict where and when crimes would occur, and then send police to those areas. The algorithm would also be able to direct police to specific people, based on their likelihood to commit a crime. Besides the obvious implications of overarching govenmental power and the potential for abuse, there's also the issue of fairness, especially in the case of impoverished communities and individuals. This is a direct violation of the principle of fairness, as it would be targeting a specific group of people, and would be using AI to do so. These kinds of systems are designed, in theory, to target the people that are most likely to commit a crime, but in reality, they target the people that are most likely to be arrested. This also creates something called a feedback loop, where the people that are arrested are then more likely to be arrested again, and so on. These algorithms would also need to breach the privacy of the people that they're targeting, which is another issue that this presents. On top of this, there's also the issue of accusing someone of a crime that they haven't committed yet. George Orwell's 1984 is a great example of this, where the government is able to arrest people for what he defines as thought crime - crimes that they haven't committed yet, but that they are "likely" to commit. All of these together create a system that is inherently unfair to the people that it's targeting in his book. There's also no transparency or accountability in a system that implements this kind of policing, which ties in with the AI that would run it. The AI would be unable to explain itself, and would therefore be unable to be held accountable for its actions. This is a violation of the principles of accountability, transpraeency, and fairness all at once.

4 Conclusion

Overall, AI is a technology that is still in its infancy. It's still being developed and improved upon. It's shaking up the world of technology, and it's changing the way that we think about the world. However, it's also creating a lot of problems that we need to address. We need to be able to hold AI accountable for its actions, and we need to be able to explain how it works, in a way that everyone can understand. We need to be able to make sure that AI is fair, and that it's not being used to target specific groups of people. In general, we need to be able to ensure that AI is being used responsibly, and that it will not and cannot be used to harm people. For right now, AI is a fun toy that we can play with, but should never be used for anything serious, especially to make decisions that are possibly life changing.

References

Maclure, J. (2021). Ai, explainability, and public reason: The argument from the limitations of the human mind. *Mind and Machines*, *31*, 421–428. https://doi.org/10.1007/s11023-021-09570-x

Stechly, M., & Kambhampati. (2023). Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv*. https://arxiv.org/abs/2310.12397v1