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 know 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

To start, what is transparency? In the general English language, 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 generally 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. To apply this principle of transparency to policy-making, companies will generally hold meetings that are accessible to the public in some way. This creates the ability for people that aren't direcly 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.

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 just says what it's going to say and doesn't think twice about any of it. 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. Some people (with far too much time on their hands) have convinced AI models such as the Snapchat AI that $2 + 2 = 5$, which, as far as I know, is very much wrong. The issue is that we don't 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, 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 easily and quickly, and generally 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 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. In fact, many people are far more paranoid about their digital footprint than their real world footprint. AI can also tie into this. Large language models (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

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.

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.

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. ¡insert evidence from some article here¿

3 Fairness

Fairness is a bit of a tricky concept to define. There's no perfect definition of fairness that appears anywhere. As far as I know, there's also no perfect example of fairness out there either, since even in court cases, which are as "fair" as we can get them, still have a possibly biased judge giving the rulings. So to define the principle of fairness demands more than simply giving the definition of a word. With that being said, the Canadian Journal of Philosophy defines fairness as such:

> ”...if a number of people are producing a public good that we benefit from, then it is not morally acceptable to free ride on their backs, enjoying the benefits without paying the costs.”

Fairness in AI is even still not so easily summarized. Generally, we interpret fairness as having no bias. Just like I explained earlier (to which I'm debating including that paragraph) AI can be excessively biased towards any given group or individual. This is, by definition, a violation of the principle of fairness. Past that, there are some AI models, such as the Snapchat AI, that partially accept suggestions for what the correct answer is for their response. This is how people have convinced some models that $2 + 2 = 5$. With that being said, most models don't accept these suggestions, simply because then their data will become more and more biased as time goes on and more users utilize that interface. For the ones that do however, they become very biased very quickly, as this can be comparable to feeding the AI new data. This data, by the nature of it being unchecked and created by anyone on the internet, can be anything, biased or not.