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