Review Essay

Lucas Vas

In her article, "AI, Explainability, and Public Reason: The Argument from the Limitations of the Human Mind," Maclure argues not that the use of AI is inherently bad, but that there are some serious ethical concerns that need to be addressed about the explainability of every AI before it's used to make proper decisions. She argues that the use of AI in the medical field and the criminal justice system are two of the most important areas where explainability needs to be a high priority. She primarily creates her argument by using the ethical theory of public reason, which is the idea that all decisions made by the government should be able to be explained to the public in a way that they can understand. She argues that the use of AI in these two fields is a violation of public reason because the AI is not able to be explained in a way that the public can understand. I would agree with her assessment, however I also think that sometimes there are good applications of AI that don't always necessarily need to be fully understood.

When AI came about in mainstream (programmers have been hard at work on it for decades) it was immediately seen as a huge time saver. Before AI truly became usable in the real world, there were movies like *Terminator* and *The Matrix* that attempted to warn people of AI taking over their lives. While I believe that this is still a distinct possibility, I believe that the more immediate concern, or maybe even a leadup to those events, is that AI, as of yet, cannot be properly explained. This creates a whole host of problems, especially since we are already using this technology in fields where explainability is a legal requirement. The use of this AI as a major decision-maker implies that the AI is able to make decisions that are better than humans, but if we can't explain how the AI is making this decision, can we really be sure that it's true? Many times, we also don't have the opportunity to challenge the decision that the AI makes, creating a situation where we're forced to

either trust the AI completely, or disregard it completely.

Maclure's argument is that the use of AI in certain fields, specifically the medical and criminal justice fields, is a violation of public reason. She presents her argument referencing the way that these fields were working before AI was introduced. She says this about the medical field:

> "For example, if IBM's Watson Health makes a surprising diagnosis or recommends an unconventional treatment, the medical team has the duty to explain to the patient why the diagnosis was made or the treatment recommended. This is required by the norm of informed consent." [1]

She also says this about the financial field:

> "To take another example, if a financial institution uses a predictive algorithm to decide whether a person qualifies for a loan, the applicant must know that it's because of some specific facts about her personal finance or credit history and not because her gender or skin colour that she was turned down." [1]

As we can see, understanding the way the AI works is quite important. Right now, if your AI is racist, you would never know. To take an extreme example combining both of these, if an AI was used to diagnose you with a severe disease that needed to be treated with an extremely expensive surgery, and you were denied a loan due to a decision made by an AI, then you would never know why. The people that used the AI could never create a proper explanation as to why you were diagnosed with this, or even if that decision was correct, not to mention whether or not you were denied the loan because of factors that are completely out of your control. This is a blatant violation of the principal of public reason.

Maclure's argument uses public reason almost exclusively, however looking at this issue from other angles can also create some interesting results, and can help to strengthen her argument. For example, if we use the ethical theory of utilitarianism, there are still violations. Creating a system that is not explainable to the public creates a possibility of harm to the general public. If we use the example of the medical field again, if an AI is used to create this diagnosis again, and it's wrong,

then the patient could be seriously harmed. If the patient is harmed, then the AI has failed to do its job, and has caused harm to the patient. This is a textbook violation of utilitarianism. Since there is no way of challenging this result, then there is no way to prevent this harm from happening, both in the initial case and the future cases. This is a serious problem, and one that needs to be addressed.

Another ethical theory that could be used for this is the theory of virtue ethics. This theory is based on the idea that the best way to live your life is to live it in a way that follows some sort of role model. I think that there are two different ways to look at this theory in relation to our examples. The first is that the AI itself is the role model, and that we should be following the AI's decisions to the letter. The second is that the AI is following a role model itself, and that its decisions are based on someone or some*thing* else. This argument also uses the virtue of universalizability, which is defined as such by Kant, then explained by Shafer-Landau:

> "...the principle of universalizability: An act is morally acceptable if, and only if, its maxim is universalizable." [2]

Using Kant's definition of universalizability, we much also understand what a maxim is, and how it is universalizable. Shafer-Landau defines a maxim as such:

> "A maxim is a principle of action that you give yourself when you are about to do something." [2]

This means that, when we are about to do something, we must ask ourselves if we would be okay with everyone else doing the same type of thing, even under different circumstances.

If we look at our first example where we use the AI as our virtuous person, then we can see that the AI may not be a good role model. If the AI is making decisions that aren't explainable, then we, as the people that are following said AI, should be unable to explain our own decisions. This creates an interesting view on how the world might work. Judges, especially in cases such as the Supreme Court, would not be required to explain their rulings and what they're based on. This would mean that anything a judge says, due to the fact that the rulings of the court are law,

would then also be law. This literally translates into "whatever the judge says is law." This would then mean that, due to the possibility of dishonesty, the decisions that a judge can make would no longer be justifiable, and could be heavily opinionated. Using the AI, which we already know can be incredibly opinionated, utilizing the AI as our role model in medical, judicial, and financial decisions (to name a few) would be a violation of virtue ethics.

If we look at our second example under the second interpretation, then we would say that the AI is following a role model. of its own. This could mean that the AI is following some sort of set of rules that it has been programmed to follow, or that the AI has found a person to rely upon. This also raises another issue - as of right now, with our explainability issue, we don't know which one of these is true. If the AI is following a set of rules, then we can't be sure that the rules are the rules that we've set for it. If the AI is following a person as its role model, then we can't be sure that the person that it's following is what we would consider a good role model. We can hope that these are true, but we can never be sure. This is another violation of virtue ethics.

Overall, the fact that the AI is not able to be explained raises a series of ethical concerns. The fact that we can't understand what it is doing in the background, nor can it explain itself to us, means that we can't be sure that it's making the right decisions, or at least the decisions that we think are right. This is a serious problem, and one that must be properly addressed before creating a system that relies on AI to make decisions. This technology is new, and being unable to understand how things work is a part of development, even in other fields. However, AI, even in its current state, is being used to create decisions that affect the lives of people. Some of the decisions may be correct, but do you really want to take that chance all the time?

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

Shafer-Landau, R. (2014). *The fundamentals of ethics, 3rd edition.*