Otto Mättas
6324363
24 March 2020

Rethinking Responsibility for AI

From Goals to Preferences

We are finding limitations to AI development not only in the physical world and computationally but also in a more philosophical sense. We keep discussing and arguing on topics like ethics and responsibility. Whether or not a machine should be held accountable. Is consciousness really biological and so on.

I argue that current views on Machine Responsibility stem from past sixty years of goal-oriented implementations and is not beneficial to either the field of study or humankind (the ultimate benefactors) as a whole. We should take a step back and rethink what we want to apply AI to and how to get best results from the application today instead of trying to close the responsibility gap. Even in the case of AGI, we could rethink how to achieve the goal, the super-intelligence, the singularity.

Where did we go wrong?

Now, what is a goal, really? Something we want the agent to achieve, that is of course given. But should an agent be able to manage its own goals autonomously? AGI is certainly expected to behave in such a manner. And that brings the question of responsibility. I think we are not really ready for that last question yet. We unreasonably expect AI to act as if human already today while we haven't figured out what being human means (e.g consciousness). So we simply cannot give a good answer yet and should not be focussing on it.

Why do we expect so much out of machines?

We have many successful weak AI solutions which can't really be scaled up to make strong AI. Unfortunately, the expectation for such opportunity is there. This has put us on an uncertain path in terms of how we develop our smart machines today. Most initiatives approach problems from a binary goal-oriented perspective. An agent might be effective at playing a game with clear winning conditions or negotiating commitments in a multi-agent environment so that the "owner" of the agent has the smallest cost of reaching the goal. This sort of an agent might not be that useful for helping a person in making their pizza. Think of all the steps that might be involved in the changing environment like preference for that day and previous preferences. How do we feasibly introduce all the necessary goals involved?

Could preferences replace goals?

So instead of pushing for machines to get better at reaching goals, we could benefit more from machines that operate on the idea of preference/feedback. In a way, this creates an implicit goal of serving the human as best as it can. This concept should be quite understandable as is quite comparable to children growing up and how parents help them reach maturity. The agent ("child") will create/update their ontology (e.g "beliefs") based on interactions with the human ("parent"). If the task is critical, ask for preferences before taking action. If the task is not that critical, allow for more learning and ask for feedback afterwards instead. This will essentially take away the problem of responsibility as the actions are directly informed and won't make for a situation open to interpretation. The human will always be responsible for their agent until it achieves maturity. For the sake of argumentation, we could say maturity is achieved when AGI is formed. This brings on the problem of maturity but takes away the unreasonable expectation towards implementations. We can then focus more on finding the common ground on what a mature agent is rather than pushing for human-like features in all AI solutions and thus inhibiting progress. Also, maturing will be an essential part of any agent so this will probably bring rapid development to the topic of maturing itself, resulting in quick resolution.

What does future look like (for humans)?

As Stuart Russell has said looking ahead - we will need to become good at being human. While most governing could be eventually done by machines, we still want to have control and responsibility towards our future. If we rethink how we develop AI technologies, we might still have a chance. If we build machines that ask for direction in uncertainty rather than rely on imperative goals, there's a good chance we will avoid even the most gruesome AI-led world-end scenario. (AI turning itself off can be preferable while it can not be a part of reaching an imperative goal). There's also a good chance that we become better ourselves along the way as teaching has proven to be a great way to learn as well.

Sources

* Human Compatible: Artificial Intelligence and the Problem of Control – October 8, 2019 by Stuart Russell

* Sharing Moral Responsibility with Robots: A Pragmatic Approach.Gordana Dodig Crnkovic & Daniel Persson - 2008 - In Holst, Per Kreuger & Peter Funk (eds.), Frontiers in Artificial Intelligence and Applications Volume 173. IOS Press Books.

* Noorman, Merel, "Computing and Moral Responsibility", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/>.

- Provided in the course:
  - Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21.
  - Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and information technology, 6(3), 175-183.
  - On the morality of artificial agents, L Floridi, JW Sanders, Minds and machines, 2004, Springer
  - Braham, M., & Van Hees, M. (2012). An anatomy of moral responsibility. Mind, 121(483), 601–634.
  - The relation between forward-looking and backward-looking responsibility. Van de Poel, I. (2011). In Moral responsibility (pp. 37–52). Springer.
  - Nagel, Thomas (1979). "Moral Luck". Mortal Questions. Cambridge: Cambridge University Press. pp. 24–38. OCLC 4135927
  - Pettit, P. (2007). Responsibility incorporated. Ethics, 117, 171–201.
  - Coeckelbergh, M. (2016). Responsibility and the moral phenomenonology of using self-driving cars. Applied Artificial Intelligence, 30(8), 748–757.