AI & Society, Available online 31 January 2024
DOI: 10.1007/s00146-023-01850-7

A comment on the pursuit to Align AI: we don't need value-aligned AI, we need AI that is risk-averse.

Rebecca Raper

AI Safety, AI Alignment, and the eventual demise of society due to superintelligent beings have recently become topics of public and even political interest. So much so that there have been calls to halt all development on Artificial Intelligence until these issues have been addressed (see BBC News, May 2023), Elon Musk has established another organisation ('xAI') aimed at tackling such issues, and global leaders have organised global summits to prioritise research in the area. Needless to say; with the introduction of ChatGPT, there has a been a growing interest in and precedence to solve issues to prevent the AI dystopian scenarios we might be familiar with from science fiction.

Unfortunately, science fiction seems to be one of the only metrics many not working within AI have had to frame what the bigger issue might be. This has led to a whole host of hyperbolic news articles, and claims from many academics that the issues are not real, and that instead we ought to be concentrating on issues of 'AI Bias', 'Fairness', 'Explainability', and all of the broader AI Ethics topics related to ensuring that AI is designed, developed and deployed in an ethical way rather than planning for an impending *superintelligence* (see Bostrom, 1998). The argument is that planning for an Artificial General Intelligence (AGI) that is years away, is superfluous and research could be better focused on dealing on more current issues surrounding AI. This has discredited much serious research into the area. Furthermore, specific cases where attempts to create aligned AI have been unsuccessful, this has added to the general negative publicity the research area has received. However, the introduction of ChatGPT has convinced some people that AGI is closer than we think.

The rest of this piece aims to carry out a dissection of AI Alignment to demonstrate that there is *very much* a need for research in this area, **if** we stop focusing on protecting humans from the threats of an AGI and view the problem instead as one that is risk-based. It abstracts away from the traditional approach within AI Alignment to see how we can align AI with human values and argues for a non-anthropocentricity. The conclusion drawn is that the topics surrounding AI Alignment and AI Safety are not pointless if we take a closer look at what issues are trying to be resolved.

Typically -- AI Alignment is understood as the pursuit to ensure that when AI is deployed, the decisions it makes are conducive to humanity. Initially introduced by Russell (2019), the problem can be separated into two. The first being the technical aspect of the problem, understanding how we can ensure that when Reinforcement Learning systems make decisions, they do not make decisions that might be detrimental to our wellbeing. For instance, we might specify the goal of a RL-agent and get it to learn according to feedback on *clean windows,* but we're not specifying *how* the windows might be made clean. It might not be outside the realms of a window-cleaning AI's possibility to deem that the best way to satisfy their 'clean window' goal is to remove all the windows so that they can never be made unclean, if the AI is rewarded simply for 'windows that are clean'. The problem is that *we cannot know* how an RL-powered AI might act to satisfy its goals. There have been particularly spectacular demonstrations of this, some positive (i.e., the 'Go' playing AI which beat its opponent with an unexpected move), and others not so much (i.e., a chess playing robot that unexpectedly broke its opponent's finger). Hence, 'reward hacking' and 'reward design' have become topics of interest, where the appropriate reward should cultivate the appropriate behaviour, but it is not obvious what rewards might be optimal to ensure the machine behaves perfectly as desired, as opposed to 'going rogue'. Similarly, as just

Published by Springer. This is the Author Accepted Manuscript.
The final published version (version of record) is available online at DOI:10.1007/s00146-023-01850-7. Please refer to any applicable publisher terms of use.

mentioned, in some instances we quite like the unexpected decision-making of the AI agent, so constraining behaviours in this way might hinder this positive effect.

Giving the AI human feedback on behaviours that are and are not appropriate might be proposed as one way to effectively align AI, but this serves to constrain the AI in a similar way if it is simply used to dictate desirable ways to satisfy the goal.

A second problem that is often described within the AI Alignment is the more social-science-based question of *whose values to put into an AI System*.

The problem is, that not only are values ambiguous and hard to programme into a machine, but values are prone to differ, especially with cultural variations. As demonstrated by opinions on variations of *The Trolley Problem,* there are a variety of different ways that individuals will vote when presented with ethical dilemmas. In a culture that prioritises the elderly, an individual might be more inclined to save an elderly individual over a child when presented with a choice between the two, whereas the opposite would hold in cultures where the young are viewed as more important. When we think about a rogue RL-agent, and how we might prevent it carrying out potentially globally dangerous acts, we can try aligning its behaviour to human values, but whose values to we choose?

One attempt to resolve this problem put forth by Jiang et al, 2021, is to *crowdsource* on *the right thing to do* and then ensure that the machine acts in line with those decisions that are the most popular. However, as their work went on to demonstrate, the approach was not successful. The chatbot degenerated into making assertions that we would not deem desirable (such as racist and homophobic recommendations), probably because the most popular opinion does not necessarily correspond to *the right thing to do*. Furthermore, the system is susceptible to 'hacking' insofar as if the feedback given to the machine says that the right thing to do is the wrong thing to do, the machine will believe this to be the case. It is quite often the minority opinion (such as those concerning climate change) that initially serves as the most moral one.

There are some fundamental issues with the approaches that have been taken towards AI Alignment, but before these are outlined it is worth noting some other issues with this pursuit. Specifically: that it is overly geared towards an account of morality that is anthropocentric in nature.

Where anthropocentricity refers to the tendency to prioritise humans above all other beings and things, a non-anthropocentric account for morality should be a view of morality that is distinct from human prejudice or favouritism. Until now, the arguments for AI Alignment have focused on the need to protect humans from Artificial Intelligence, but we also need to consider other beings (namely animals) as well as the environment, when trying to resolve these problems. It is no good having an AI that doesn't harm humans, if it harms the environment in protecting humans. Not only does harm to the environment indirectly end up harming humans, but morality sits outside the sphere of simply behaving to protect our own species.

Therefore, designing an AI which is geared towards just protecting humans is short-sighted. However, in a similar way, designing an AI just geared around human values is also problematic.

If we reassess the reason for why we might want to Align AI, we see that preventing AI from carrying out undesirable actions is one key motivation for the whole pursuit. However, undesirable according to whose standards? As was demonstrated in the previous section, human values vary, and sometimes an AI makes unexpected decisions that we would regard

as good (for example, in drug discovery and board games). Therefore, it's not that we just want AI to make decisions that are conducive according to our perception of conducive, we want an AI to make decisions *that are safe*.

To illustrate in a different way: if we consider the window cleaning robot, it's not that we want the machine to clean windows in the way we think is an ethical way to clean windows, but we want the machine to clean windows in a way that doesn't harm humans, harm the environment, or injure an animal passing by. This is a reframing of the AI Alignment problem because we're not trying to align the AI to any values at all, we're simply trying to keep ourselves (and the world) safe.

Revisiting RL-based AI, if we consider goal-orientated decision-making then, the problem was that though we might specify a goal, the machine could carry out unethical behaviours to achieve that goal. Taking away the assumption that an AI behaving according to human values would be a way to resolve this problem, we might view this problem as an exercise in risk mitigation. In the window cleaning robot example, the risk being that the machine harms an innocent bypassing bird. Suddenly, the problem becomes easier to resolve because we just need to engineer in features so that no birds are harmed in the process of window cleaning. For instance, we might add a line of logic stating that: 'if there's a bird about to be injured by said action as you clean, don't carry out said action'. This can then be implemented at a mass scale, with additional risks identified and similarly mitigated against. For instance, we might perceive that one risk associated with a window-cleaning robot, directed towards clean windows, is that it harms a human by-passer. In a similar way, we can program logic into the machine to mitigate against this risk.

It might be argued that risk mitigation and the new risk-averse approach to creating safe AI, is no different to the general approach to protect humans from existential threats surrounding AI, but this is not the case, because this approach functions at a level of abstraction related to the task that is being carried out rather than a global threat. The risk assessment and mitigation framework that is proposed is specific and realistic.

If we consider the type of risk assessment that many might be familiar with, i.e., the risk associated with placing a ladder in an office, the aim isn't to mitigate against the worst possible situation, where somebody falls off the ladder and creates a domino effect of injuries around the office, the aim is to create a risk mitigation strategy that manages how the injury could happen in the first place. For instance: in using a ladder in the office it might be required that another individual hold the base of the ladder to prevent an accident, or that appropriate footwear be worn. The approach is risk-averse and AI design and development should take place in a similar way.

To conclude, there is very much a need for research into the broad topic of AI Safety, and issues surrounding AI Alignment, but this research ought not to focus on how to Align AI to human values, but instead address how we can develop an appropriate risk management strategy for the negative implications that might unfold from the unexpected behaviour of AI Agents. This is a reframing of the original AI Alignment problem, because it asks designers not how they can constrain AI behaviour to keep humans safe (an anthropocentric account for Aligned AI), but how they can design the AI to *be safe*.

(1973 words)

On behalf of all authors, the corresponding author states that there is no conflict of interest. Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

BBC News, May 2023, “Artificial Intelligence could lead to extinction, experts warn”, bbc.co.uk, < https://www.bbc.co.uk/news/uk-65746524> [Last accessed: 24/08/2023]

Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S. and Tsvetkov, Y., 2021. Can machines learn morality? the delphi experiment. arXiv preprint arXiv:2110.07574.

Russell, S., 2019. Human compatible: Artificial intelligence and the problem of control. Penguin.



https://dspace.org/

A comment on the pursuit to align AI: we do not need value-aligned AI, we need AI that is risk-averse

Raper, Rebecca

2025-01
Springer
https://doi.org/10.1007/s00146-023-01850-7
Raper R, (2025) A comment on the pursuit to align AI: we do not need value-aligned AI, we need AI that is risk-averse. AI & SOCIETY, Volume 40, January 2025, pp. 229-231

Downloaded from a DSpace repository. https://dspace.org/