Shashwat Vidhu Sher

MLA 373

Professor Kochenderfer

Paper I

29 April 2023

Humans and AI: A Case for Harmonious Coexistence

A big part of the research in AI is concentrated in the field of AI Safety. Boxing methods, stunting and tripwires are some of the methods proposed for the control problem (Bostrom 129-137). At the same time, designing appropriate motivational selection methods and value loading problem are active areas of research (Bostrom 143, 207). It won’t be incorrect to say that many researchers in the field are working towards developing artificial moral agents whether directly or indirectly (Wallach and Allen 33-36). It is in the service of this goal a paradigm shift is required in our mentality to ensure that we develop AI which is aligned with human values and not some arbitrary understanding of the world. In the paper I would argue for two goals – treating AI development as a special case of raising children and giving AI systems enough autonomy so that they could work with humanity as partners and not as servants.

To create moral agents that understand morality as we do, it is important that the training process is as closely aligned as it is for humans. When we talk about training humans so that they could become contributing members of society, we specially talk about imparting good education and values to our children. The case for AI should not be much different from this approach. Disciplining children in a healthy manner when they are wrong and encouraging them appropriately when they do something right ensures that they understand the consequences of their action (Unicef). Too strict a parenting is detrimental to a child’s mental growth. Similarly, if a child is rewarded for something which is supposed to be his/her responsibility, he/she could develop a sense of entitlement (PU). Taking cue from these cases, researchers should approach AI development where they do not severely ‘punish’ a system by stunting its learning process in ways that renders it cognitively underdeveloped. Also, they should not base their motivation methods on rewarding the AI for tasks that it is supposed to do. An important thing to consider is to let the AI search for a life-goal on its own. It can be guided on its journey so that its goal is not anti to that of humanity but feeding an ‘artificial’ goal can result in a system with a hidden tendency towards subversion or even a full-blown existential crisis (Maxwell 4). Only when we stop seeing AI as a tool and more as our child, we can have a higher degree of confidence in it developing the same understanding of the world as we do.

One can argue on both the points raised above about training and motivation, but the suggestions are grounded in time-tested methods that have proved efficacious. Even after the best of our intentions and teaching, it is not guaranteed that children develop into adults who are cognizant about the problems of the world and have an empathy towards fellow humans. However, the constant process of learning from the mistakes of a previous generation and correcting them in raising the next generation has essentially been the story of societal evolution. For AI development, this iterative process is our best shot in developing an intelligence which is more human-like rather than machine-like (Krishna et. al.). On the problem of motivation, one can ask if an AI system that is allowed to come up with its own goals could really be useful for humanity? The answer again lies in the way we bring up our children. The idea is that with human-like values, an AI system will be responsible enough towards society and use its capabilities for the welfare of general public. Simultaneously, it would be free to pursue a goal that aligns with the AI’s understanding of being happy and fulfilled. A human-like intelligence can’t be achieved without providing a human-like autonomy (Oudeyer 1). Hence, a similar approach to AI training and motivation as we have for humans, is not irrational to pursue.

Once a system develops enough intelligence, the immediate next question is about respecting the autonomy of the system for its full cooperation. This is where a critique of capability control methods can be very useful. Any human-like intelligence can understand being restricted by methods like boxing, stunting and tripwires. These techniques are essentially equivalent of keeping someone as hostage while trying to have them do your bidding. It is not far-fetched to conceive that such an entity would develop a resentment for its ‘captors’ and would search for ways to escape and possibly even retribution. Cooperation and harmony can only be expected if we respect the rights of other party (Parikh). Practical concerns for the misuse of the tremendous powers an AI would possess are valid but methods to keep those powers in check should not be based on physical and mental harm but sanctions and non-cooperation (Balliet et. al. 609-610). The more an AI system can be designed or trained to have human-like values, the more would be its need to depend on humans and work with them to fulfill its own goals. Moreover, using another intelligence in a manner that is disrespectful and undermines its autonomy raises moral concerns. We don’t have to look too far into the past to understand the problems caused by slavery and its effect on the disenfranchised section of society. With the great advantage of using AI capabilities for human benefit, comes great responsibility for its ethical treatment and no trust can be built between two sentient beings without respect for each other’s rights and freedom.

For an AI system to be useful for humans, we have to treat it as an extension of humanity rather than just another piece of technology only meant to serve

UNICEF - <https://www.unicef.org/parenting/child-care/how-discipline-your-child-smart-and-healthy-way>

PU - <https://onlinegrad.pepperdine.edu/blog/entitled-children-improving-behavior/>

Maxwell - <https://openaccess.wgtn.ac.nz/articles/journal_contribution/_What_is_my_purpose_Artificial_Sentience_Having_an_Existential_Crisis_in_Rick_and_Morty/14721459>

Parker - <https://medium.com/data-driven-fiction/ai-and-existentialism-14ab95d0a236>

Krishna - <https://www.pnas.org/doi/10.1073/pnas.2115730119>

Oudeyer - <https://arxiv.org/ftp/arxiv/papers/1712/1712.01626.pdf>

Balliet - <https://psycnet.apa.org/doiLanding?doi=10.1037%2Fa0023489>

Parikh - <https://avasant.com/report/artificial-intelligence-should-robots-have-rights/>

Works Cited

Hand, David J, and Shakeel Khan. “Validating and Verifying AI Systems.” *Patterns (New York, N.Y.)*vol. 1,3 100037. 12 Jun. 2020, doi:10.1016/j.patter.2020.100037. Accessed 16 Apr. 2023.

Larkin, Zoe. “AI Bias - What Is It and How to Avoid It?” *levity.ai,* 16 Nov. 2022, [levity.ai/blog/ai-bias-how-to-avoid](https://levity.ai/blog/ai-bias-how-to-avoid). Accessed 16 Apr. 2023.

Pullum, Laura. “Verification and Validation of Systems in Which AI is a Key Element.” *sebokwiki.org*, 31 Oct. 2022, [sebokwiki.org/wiki/ Verification\_and\_Validation \_of\_Systems \_in\_Which\_AI\_is\_a\_Key\_Element](https://sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element). Accessed 16 Apr. 2023.

Schwartz, Reva, et al. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology, 15 Mar. 2022, doi:[10.6028/NIST.SP.1270](https://doi.org/10.6028/NIST.SP.1270). Accessed 16 Apr. 2023.

Silberg, Jake and James Manyika. “Tackling bias in artificial intelligence (and in humans).” *mckinsey.com*, 6 Jun. 2019, [www.mckinsey.com/featured-insights/artificial- intelligence/tackling-bias-in-artificial-intelligence-and-in-humans](http://www.mckinsey.com/featured-insights/artificial-%09intelligence/tackling-bias-in-artificial-intelligence-and-in-humans). Accessed 16 Apr. 2023.

Szegedy, Christian, et al. *Intriguing Properties of Neural Networks*. arXiv.org, 19 Feb. 2014, [arxiv.org/abs/1312.6199](http://arxiv.org/abs/1312.6199). Accessed 16 Apr. 2023.