Valuable AI versus AI with values

Ferdous Alam

The landscape

Many academics and researchers posit that the emergence of human-level artificial intelligence will be achieved within the next two decades. In 2009, 21 AI (Artificial Intelligence) experts participating in AGI-09 experts believe AGI(Artificial Intelligence) will occur around 2050, and plausibly sooner. In 2012/2013, Vincent C. Muller, the president of the European Association for Cognitive Systems, and Nick Bostrom from the University of Oxford, conducted a survey of AI researcher where 60% responded that AGI likely to happen before 2040. In 2017 May, 352 AI experts who published at the 2015 NIPS and ICML conferences were surveyed resulting in estimate that there's a 50% chance that AGI will occur until 2060. In 2019, 32 AI experts participated in a survey on AGI timing with 45% of respondents predict a date before 2060.[1]

There is little contention or disagreement in terms of the benefits AI will provide by analyzing data, integrating information and a much faster rate than humanly possible. However, how we utilize the insights and apply them to form decision making is not an easy problem to solve. The Brooking Institute mentions that "The world is on the cusp of revolutionizing many sectors through artificial intelligence, but the way AI systems are developed need to be better understood due to the major implications these technologies will have for society as a whole" [2]

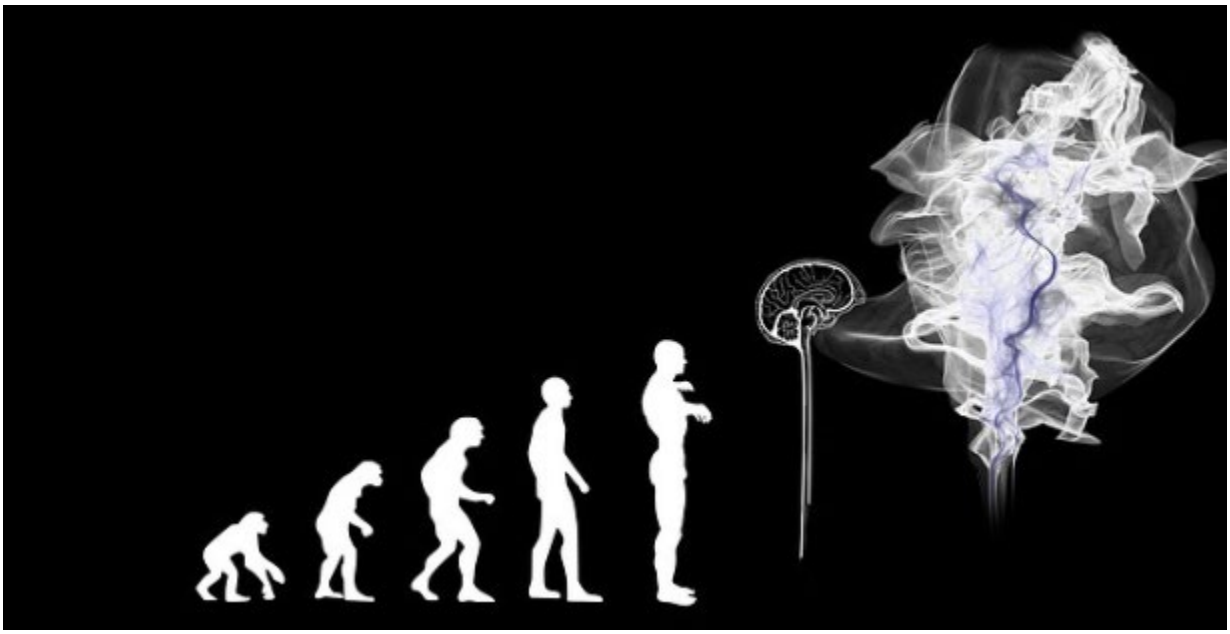
Recent surveys showed that the overwhelming majority of Americans (82%) believe that robots and/or AI should be carefully managed. This figure is comparable to with survey results from EU respondents.[3] There however is a caveat when it comes to alignment of the survey result with how we perceive or correlate intelligence with positive traits. Due to what is known as 'affect heuristic bias's we often rely on our emotions, rather than concrete information, when making decisions. This

leads us to overwhelmingly associate intelligence with positive rather than negative traits or intuitively conclude that those with more intelligence possess other positive traits to a greater extent. Hence, even though we may show an overall concern there is a possibility that we may fall into the pitfalls of miscalculating possible cost associated with AI/AGI adoption.

Embedding values

S. Matthew Liao "argues that human-level AI and superintelligent systems can be assured to be safe and beneficial only if they embody something like virtue or moral character and that virtue embodiment is a more appropriate long-term goal for AI safety research than value alignment." [4]

In his 1942, before the term AI/AGI was coined, science fiction writer Isaac Asimov in his short story "Runaround", proposed three laws or *robotics* which can be seen as a corollary that can be applied to AI/AGI. According to his proposal the First Law states: A robot may not injure a human being or, through inaction, allow a human being to come to harm. The Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Finally, the Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



While this is a novel attempt but embedding virtue or principles/laws from a consequentialist perspectives might fall short. It is argued "that it is impossible to precisely and consistently predict what specific actions a smarter-than-human intelligent system will take to achieve its objectives, even if we know the terminal goals of the system." [5]

Another heuristic approach might be to consider the four ethical principles by EU High-Level Expert Group on AI that closely resembles the commonly accepted principles of bio ethics, excerpted from Beauchamp and Childress (2008), which include: Principle of respect for autonomy, Principle of beneficence, Principle of nonmaleficence, Principle of justice.

The proposed four principles by this group when it comes to AI [6] are:

I) Respect for human autonomy - AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. This essentially seems to cover Principle of respect for autonomy, Principle of beneficence from the Bioethics

II) Prevention of harm (Principle of nonmaleficence)- AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. 30 This entails the protection of human dignity as well as mental and physical integrity.

III) Fairness (Principle of justice)–The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization

IV) Explicability -This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible –explainable to those directly and indirectly affected.

The principle of explicability seems to be a completely new addition to the previous framework which has significant implications. According to Luciano Floridi and Josh Cowls, "the addition of the principle of 'explicability,' incorporating both the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question 'who is responsible for the way it works?'), is the crucial missing piece of the AI ethics jigsaw."

Conclusion

The tradeoff between the value that AI promises versus the values we need to embed within the decision-making process is both intriguing and challenging. Moral values and principles in terms of systematically investigating what makes acts right or wrong has been debated for eons. While an objective value system that is optimal is unlikely to emerge anytime soon, yet the various perspective and the framework proposed would serve as a starting point which we can use to look at different perspective and strive towards a better solution.

References:

1. Cem Dilmegani, (2022). When will singularity happen? 995 experts' opinions on AGI. https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/

2. Darrell M. West and John R. AllenTuesday, (2018). How artificial intelligence is transforming the world. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

3. Baobao Zhang and Allan Dafoe (2019) Artificial Intelligence: American Attitudes and Trends https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/executive-summary.html

4. S. Matthew Liao (2020). Ethics of Artificial Intelligence https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190905033.001.0001/oso-9780190905033-chapter-14

5. Roman Yampolskiy(2019). Unpredictability of AI https://www.researchgate.net/publication/333505954_Unpredictability_of_AI

6. Independent High-Level Expert Group on Artificial Intelligence (2019) https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

7. Luciano Floridi and Josh Cowls, (2019). A Unified Framework of Five Principles for AI in Society https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/7