Importance of Ethics in Artificial Intelligence

Baanee Singh

January 30, 2025

Trust and Ethics for Autonomous Vehicles

After reading the paper, "Perspectives on Ethics of AI: Computer Science", by Ben Kuipers, in Markus Dubber, Frank Pasquale, and Sunit Das (Eds.), we see the role of ethics in artificial intelligence (AI), especially in the development and usage of autonomous vehicles (AVs). For AVs to be accepted on the road, humans need to be able to trust their behavior as explained in the paper. This includes adhering to ethical principles and social norms that ensure the safety and behavior of vehicles in real world situations. The main argument is that trust in AVs can be built through adherence to social norms such as not deliberately harming humans, having competent driving skills similar to skilled human drivers, and actively avoiding dangerous scenarios.

I agree with the paper's perspective on the importance of ethics in AI. AVs operate in human environments and must meet high ethical and safety standards to ensure public trust and support. Integrating ethics into AI development demonstrates the necessary responsibility and foresight to address potential risks and social concerns. Without this vital integration, AVs risk public rejection due to fear and mistrust.

The paper discusses the example of the Deadly Dilemma, also known as the Trolley Problem. Here, we can highlight the specific social norms and ethics that AVs must adhere to in order not only to be useful, but also to be trusted by humans. I largely agree with the social norms presented in the paper, as they are the basic ethics that should be adhered to by AVs and AI as a whole. The first one (SN-1) presented is that "a robot will never deliberately harm a human being."[1] This aligns with the basic ethical principles, such as Asimov's First Law of Robotics, described in the paper.

It provides a clear moral guideline. However, this does not mean that it will eliminate all harm or that it could never occur, which is a concern raised. Second, we find that ”in a given situation, a robot will be no more likely than a skilled and alert human to accidentally harm a human being”[1] (SN-2). This norm is a reasonable standard for competence, but also the bare minimum when comparing AVs to skilled human drivers. It sets the bar for performance while also understanding the limitations of human and machine decision making. The third norm (SN-3) is that ”a robot must learn to anticipate and avoid Deadly Dilemmas.”[1] This norm encourages AVs to learn and adapt to avoid dangerous situations as lives could be at stake. It is similar to how we expect generative AI to work, by providing it information or a question that it could answer and then adapting the answer or providing a new answer to other follow-ups asked by the user. Counterfactual thinking and continuous learning are essential for building safer AVs and AI systems.

I would propose adding another social norm that focuses on transparency and communication between AVs and humans. This way the AV would be able to communicate its intentions to humans when they are interacting with the machine. An example here would be if an AV plans to stop at a crosswalk where it senses a pedestrian walking; it should be able to signal this not only to the pedestrian but also to the person sitting inside the vehicle. Since there is clear communication and transparency between the AV and humans, trust will be built in the ethical behavior exhibited by the AV that makes humans trust the system and the safety it will provide.

References

[1] B. Kuipers, “Perspectives on Ethics of AI: Computer Science,” *Oxford Handbook of Ethics of AI, Oxford University Press, 2020*, Aug. 2019.