Talk concept and references

Benjamin Kahl
Seminar: *Normative Reasoning and Machine Ethics*
Talk: *AI Systems and Bias and Unfairness*
Primary References: *[5], [2] and [6]*

The underlying goal of the talk is the examination of demonstrable human bias towards AI systems, more specifically, autonomous vehicles and the consequences implied by it.

The *Center for Humans and Machines* at the *Max-Planck Institute for Human Development* has conducted several experiments in this regard. Thus, their publication catalogue serves as the primary source for references.

The talk will be structured into the following three parts:

First part

The first will examine general opinions and perspectives on how autonomous vehicles *should* behave when confronted with trolley-problem like scenarios.

How Should Autonomous Cars Drive? A Preference for Defaults in Moral Judgments Under Risk and Uncertainty [5] by Meder et al. investigates this conundrum in a series of two empirical studies that portray approximations of real world scenarios to test subjects and extrapolate what is perceived as morally acceptable behavior for AVs under the following circumstances:

- When the consequences of an action or inaction are only probabilistically known or entirely unknown.
- When the consequences of an action are fully known in advance (retrospective evaluation).

Furthermore, *The Moral Machine Experiment* [1] by Awad et al. may also provide useful data for this segment, but ought to be omitted, due to collisions with a different, subsequent talk "The Moral Machine Experiment and its Consequences".

Second part

With a reasonable baseline established as to how AVs *should* behave, the second part will be dedicated to *how* people believe AVs will (or are) behaving. Contrasting this outlook with the results from the previous part will satisfy the primary focus point of human bias towards AVs.

Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation [2] examines how blame is distributed between human and machine drivers when either one or both drivers make a mistake.

Third part

Lastly, *The social dilemma of autonomous vehicles* [4] and *Psychological roadblocks to the adoption of self-driving vehicles* [6] may serve as an addendum two the second part, as both analyze the contradictory preferences of consumers on whether if a car should act in a purely utilitarian fashion or prioritize its own driver.

Should it be found that the talk substantially overlaps or repeats material from subsequent talks “Autonomous Vehicles: Overview and Ethical Challenges” and “The Moral Machine Experiment and its Consequences”, the primary focus point can be respectively adjusted.

Given that all of the above listed publications mainly rely on scenarios similar to the trolley-problem, it may be beneficial to briefly examine this models shortcomings on the basis of *The Trolley, The Bull Bar, and Why Engineers Should Care About The Ethics of Autonomous Cars* [3].

References

1. Dsouza S. Kim R. Schulz J. Henrich J. Shariff A. et al. Awad, E. The moral machine experiment. *Nature*, 2018.

2. Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B. Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation, 2018.

3. J. Bonnefon, A. Shariff, and I. Rahwan. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, 107(3):502–504, March 2019.

4. Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

5. Björn Meder, Nadine Fleischhut, Nina-Carolin Krumnau, and Michael Waldmann. How should autonomous cars drive? a preference for defaults in moral judgments under risk and uncertainty. *Risk Analysis*, 08 2018.

6. Bonnefon J. Shariff, A. and I. Rahwan. Psychological roadblocks to the adoption of self-driving vehicles. *Nature*, 2017.