Talk structure

Benjamin Kahl
Seminar: *Normative Reasoning and Machine Ethics*
Talk: *AI Systems and Bias and Unfairness*

Due to potential collisions with other talks as well as a deeper understanding of the underlying publications the proposed talk structure slightly differs from the originally proposed *talk concept and references*.

The talk will center around human biases and fallacies which may prevent or hamstring the widespread adoption of autonomous vehicles and how to combat these.

Introduction

To initiate, a brief overview of the current status-quo in regards to AVs will be presented in addition to some brief data reflecting general opinions on AVs.

Roadblocks

Nature magazine article *Psychological roadblocks to the adoption of self-driving vehicles* [4] outlines a total of three psychological roadblocks that prevent a widespread acceptance of AVs.

The bulk of the talk will introduce each one and subsequently dissect it by presenting an independent, corresponding study which may either confirm or refute the claimed roadblock as well as provide further information on potential counter-measures and consequences.

Each roadblock and its corresponding study will be presented in the following order:

- **Roadblock 1:** *Dilemma of autonomous ethics*
  Referring to the general consumer preference to AVs which prioritize the safety of their own driver over pedestrians which run in direct opposition to the common mindset that AVs ought to minimize accident casualties.
  The studies included in Science magazine article *the social dilemma of autonomous vehicles* [2] confirm the prevalence of this dilemma.
  More importantly, the authors also deconstruct some of the commonly proposed counter measures, such as government-imposed regulations, and propose some solutions of their own.
- **Roadblock 2:** *Overreactions to inevitable accidents*
  Stigma produced by inevitable AV traffic accidents may be unfairly biased towards the AV industry due its novelty.
  However, the experiments conducted in *Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation* [1] seem to suggest the opposite: Human drivers being blamed more often than their machine counterparts in the case of accidents being caused by both.
  The results of the studies will be presented as well as the conclusions drawn by Awad et al. to examine in which specific scenarios the above described roadblock applies and in which it does not.
- **Roadblock 3:** *Lack of transparency in machine decision making process*
  The final roadblock listed by Bonnefon et al. is the lack of trust in AVs caused by the lack of transparency into the underlying decision making process.

Interestingly, the findings made in *How Should Autonomous Cars Drive? A Preference for Defaults in Moral Judgments Under Risk and Uncertainty* [3] suggest that, despite Roadblock 1, a purely utilitarian ethics model is not always seen as the most moral.
The authors argue that adopting a *rule-consequentialist* model may be appropriate for most *real-world scenarios*, implying that the above described roadblock exists purely as a consequence of often invoked *hypothetical scenarios* such as the trolley problem.

Conclusions

To conclude, an overview will be provided for each of the three roadblocks.

References

1. 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.

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

3. 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.

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