Non-intervention policy for autonomous cars in a trolley dilemma scenario

Bianca Helena Ximenes (Informatics Department, Universidade Federal de Pernambuco; bhxmm@cin.ufpe.br)

DOI: 10.1145/3236644.3236654

Introduction

This column presents early work on human biases and preferences described in research from Philosophical and Social Sciences, and discusses their impact in AI ethics, especially concerning autonomous vehicles.

Human nature and decision biases

In many contexts, humans prefer to withdraw their participation in scenarios where decision is too complex or have convoluted ethical implications. Often, they will let chance, or time, or another exogenous factor force a decision, so they are spared the choice and its consequences. Surely, that effectively means making a choice in the end; but not opting explicitly makes it easier for humans to deal with their own conscience.

By recognizing such effects, the present column aims to discuss the following problem: *Is a non-intervention policy in trolley dilemma scenarios a desirable way for humans to interact with autonomous vehicles?*

The discussion that follows is based on the premise that practitioners responsible for autonomous vehicles have a moral obligation of ensuring their full functionality to the best of their ability, and that saving lives is a golden rule. However, it is also based on the premise that scenarios such as that of the trolley dilemma will unfortunately be present and understanding human limitations and preferences might prove useful to modeling.

Killing or letting die?

One of the paramount aspects of trying to establish a rank of societal priorities from human research or questionnaires is the framing effect. Seminal research, such as the one that granted the Nobel of Economic Science to Daniel Kahneman and Amos Tversky shows how the same two sets of options, framed differently, yielded a completely different final collective preference concerning what to do in a critical scenario of an epidemic. Moreover, one of the factors that drove the change in the volunteers opinion was the use of the word kill, which is negatively charged and directly related to trolley dilemma scenarios (Tversky & Kahneman, 1974) (Tversky & Kahneman, 1981).

One of the earliest discussions of the trolley dilemma itself had Thomson argue that some people found that letting die had a different moral weight than killing, and different choices ensued according to such perception (Thomson, 1976). She followed presenting multiple dilemmas, each framed with a slight difference from previous others, and the results of what was deemed admissible or otherwise varied according to each factor. Often these factors were information of how the scenario came to be, and relied on extensive background information, such as the person in the tracks was a child, the person was there illegally and knowingly, the person was put there by a villain, or the person was randomly assigned to be there. These scenarios brought other papers that discussed each nuance more detailedly. This is summarized in Figure 1.

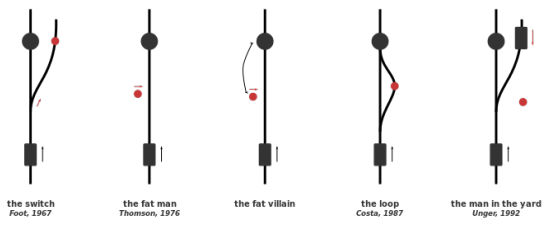

Figure 1: Trolley dilemma scenarios of Foot (1967), Thomson (1976), Costa (1987), and Unger (1992)

None of these nuances will be taken into account by an autonomous vehicle, as such information will not be available at the time of an accident. Therefore, it is not possible to rely on human objectivity when establishing an immutable ranking as a parameter to an algorithm.

Finally, in the more recent article (Waldmann & Dieterich, 2007), authors reframe once again the trolley problem and propose a simulation experiment. They find objectively that the moral standards fluctuate according to different framing of scenarios and background information, and classify it as an intervention myopia. Hence, non-intervention might be a more viable path.

A universal guideline for saving lives

The MIT Moral Machine (*MIT Moral Machine*, 2016) makes it clear that humans value actions and lives differently. Figure 2 depicts the general preference on whose life or which principle should be considered as most important when deciding whether and where to divert a cars course.

Not all preferences were rated and ordered in a single rank, being instead translated into preferences between pairs of factors that could be demographical (i.e.: age, fitness, gender, species) or more related to personal belief (i.e.: avoiding intervention, individuals social value, number of lives, protecting passengers, upholding traffic laws).

By using millions of data points to train a Machine Learning model in sufficient variable scenarios, it would be possible to achieve a general complete ranking of the value of lives and actions that reflected the judgments of the majority, use it to establish rules for autonomous cars, and release them on the streets once they are ready.

Doing so is technically possible. However, choosing who to let die might be too serious a choice to be left up to personal opinion or to an algorithm, especially when it is universally applicable across different countries and cultures. It can be argued that it is neither fair nor admissible to extend the opinions of the majority to an issue of literal life or death; neither would it be ethical to rank humans' lives, especially when there are no apparent reasons

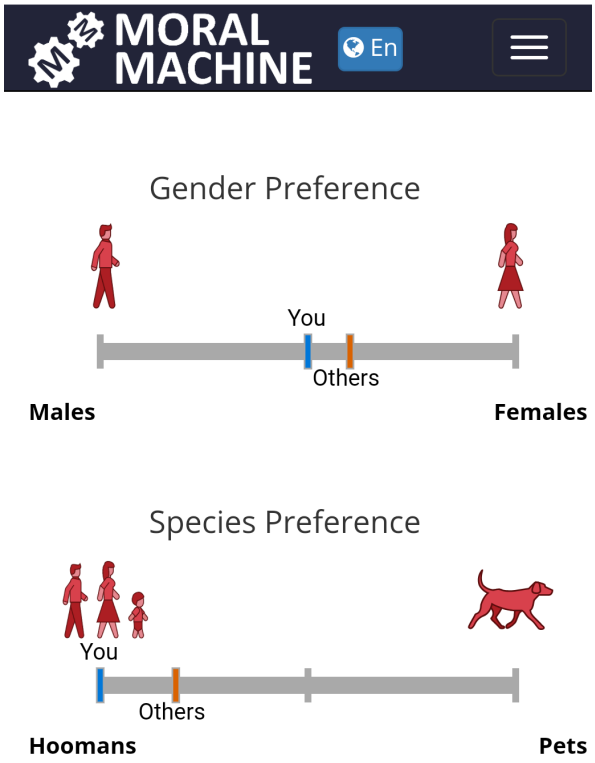

Figure 2: Partial screenshot of the result of the MIT Moral Machine judgment game. The results highlight how the judgment of the taker compares to the overall results

for specific choices other than personal preference. This scenario of automated pre-defined choice could be considered a breach of Human Rights, as one specific demographic profile would be in practice marked as inferior or less socially costly if they were to die. Considering the recent UK House of Lords' document on Artificial Intelligence and its several related subjects and implications. It states in one of its summary points that "The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence." However, by implicitly attributing value to different lives and personal characteristics, the result could be framed as the AI ultimately choosing to hurt or destroy a human being - the one who ranks lower according to the algorithm. This choice is not to be made deliberately, as it is one step further weaponizing AI to act against specific groups of people.

Another issue that comes up from Social Science research is the impossibility of carrying

out perfectly democratic votes that conform to the principles of being transitive, reflexive, and complete when there are more than two choices. This is discussed in the Impossibility Theorem by Kenneth Arrow and underpins much of the Economic Theory of choice. Arrow even concludes that the only fair system after all would be a dictatorship, even though it was not desirable (Arrow Kenneth, 1951). Arrows theorem has been addressed in research throughout the years, but it is not a solved conundrum (Frohock, 1980). As seen in the scenarios of the Moral Machine, there are well over 2 options one can choose from, which falls into a classical scenario where a democratic, fair choice is not attainable.

Choosing not to choose

The difficulty of making choices is also studied in other behavioral sciences, and go beyond cognitive biases of the human incapability of calculating outcomes into the field where humans attempt to avoid explicit choices altogether, preferring to let time run out than commit to a single option (Shin & Ariely, 2004)

Leaving the choice of who to let die to chance "or non-interventionism" may carry other results that have to be further researched but show promise. It unburdens the autonomous vehicle user both because they know the car will not explicitly choose another life over their own (in the case the algorithm is set to save pedestrians), and because they do not need to feel responsible for complying with an algorithm that ranks and weighs people's lives. The research of Kelly shows that parents who have experienced moral decisions derived from first pregnancies where the fetus had disabilities or genetic deficiencies detected while in the womb only chose to try to conceive again 34.5% of the time. On the other hand, parents who experienced first pregnancies where the fetus died for reasons other than malformation were much more prone to trying to conceive again, with over 85% opting for parenthood. Kelly concluded that the families in the former case not only had an emotional burden to carry, they also had to make ethically complex choices, such as terminating the pregnancy or deciding whether a disability was indeed good reason for an abortion. In the end, the majority of them preferred avoiding the issue altogether.

The direct repercussion for implementing a self-driving algorithm that does not take into consideration the unwillingness of humans to be faced with complex, ambiguous choices, is that autonomous cars may face resistance in adoption. The benefits of self-driving cars are such that they are being embraced and regulated by governments around the world, seen as a way to move past the most ubiquitous reason of traffic accidents and deaths: human error. An example is the SELF-DRIVE Act passed in 2017 by the United States of America House of Representatives (The Senate and House of Representatives of the United States of America, 2017). But that reason by itself might not be enough for persuading users to make the change; especially if they find that they do not agree with ranking parameters and model results, or that they will face a complex moral conundrum every time they take the car out for a drive. It is one thing to be faced with an unwanted scenario, such as a trolley problem, and make a decision in the moment. It is another to leave the house knowing which decisions have been made.

Conclusions

Non-intervention as a policy for autonomous vehicles is something hard to discuss, but it may prove a viable option due to three major questions discussed in this column: (i) the framing effect and intervention myopia, due to humans being very sensitive to changes in context; (ii) the impossibility to reach with a universal rank across nations and cultures, and how model results can be more easily diverted to uses that were unintended; and (iii) the difficulty to deal with complex moral questions when the output is known or considered too risky.

By considering these topics, we can open new frontiers on moral, ethical, and AI research. It may be hard to accept humans are not able to control these scenarios, but open discussions need to be carried out to assess whether the benefits of a non-biased system a system that relies on chance can outweigh the perils that come with what humans are building.

References

Arrow Kenneth, J. (1951). Social choice and individual values. *Cowles Foundation*.

Frohock, F. M. (1980). Rationality, morality, and impossibility theorems. *American Political Science Review*, *74*(2), 373–384.

Kelly, S. E. (2009). Choosing not to choose: reproductive responses of parents of children with genetic conditions or impairments. *Sociology of Health & Illness*, *31*(1), 81–97.

Mit moral machine. (2016). http://moralmachine.mit.edu. (Accessed: 2018-05-30)

Shin, J., & Ariely, D. (2004). Keeping doors open: The effect of unavailability on incentives to keep options viable. *Management Science*, *50*(5), 575–586.

The Senate and House of Representatives of the United States of America. (2017). *Safely ensuring lives future deployment and research in vehicle evolution act, or self drive act (hr 3388).* (https://www.congress.gov/bill/115th-congress/house-bill/3388/text)

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, *211*(4481), 453–458.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological science*, *18*(3), 247–253.

Bianca Helena Ximenes
Bianca is a cross-disciplinary researcher and Doctoral Student at Universidade Federal de Pernambuco. Her research spans Human-machine boundaries and interaction, Innovation Management, Behavioral Economics, and Digital Ethics. Her community impact earned her the Google Developer Expert in Product Strategy recognition. She is member of the IEEE P7010 Working Group and a Sci-Fi aficionada.