A Contractarian Ethical Framework for Developing Autonomous Vehicles

by Mihály Héder(MTA SZTAKI)

The way forward for autonomous vehicle ethics does not revolve around solving old moral dilemmas, but on agreeing on new rules.

Contractarian ethical frameworks claim that the norms we accept as good or proper are mere results of social compromise that is ultimately driven by the self-interest of the involved parties. This position is in contrast with other paradigms around the foundations of ethics, for instance virtues or divine commands.

We are under no obligation to subscribe to one single, exclusive ethical paradigm for all purposes and aspects of our lives. One could apply a particular approach to autonomous cars while allowing others in other domains as long as they can be made compatible.

We believe that a contractarian approach should be taken in the context of autonomous cars, and also that if we are to ever enjoy a serious diffusion of fully autonomous cars it will happen based on the grounds of compromise - or it won't happen at all.

From this it follows that the decisions required during autonomous car development are to be found at the intersection of what is generally considered to constitute acceptable vehicle behaviour as applies to all road users - if such an intersection exists at all. This means that the industry involved in defining such behaviour should simply make proposals and ask for a compromise rather than chasing for moral truths.

The case of autonomous cars should be easier than other social issues, too, because any person can conceivably take on the identity of any type of road user in a particular situation. An individual may be a pedestrian in the morning, a bicycle rider during the day and a passenger in the evening e.g. in an autonomous cab. With other issues our identities tend to be more entrenched.

Let us take most basic autonomous vehicle related ethical dilemma to illustrate the approach. The autonomous car finds itself in an emergency situation in which it can either hit and kill a group of pedestrians or swerve and sacrifice its passengers [1]. There appears to be no other option. This thought experiment has been advanced with a variety of discriminating factors like the number of casualties in pedestrians/passengers, age, gender, various forms of social role of the involved people, etc.

The example reveals the very high dependence on both our and the car's epistemic facilities in evaluating such situations.

In reality the car cannot be certain what kind of objects it has detected as the Arizona Uber incident in which a cyclist died illustrated. Worse still, it has only a partial appraisal of the

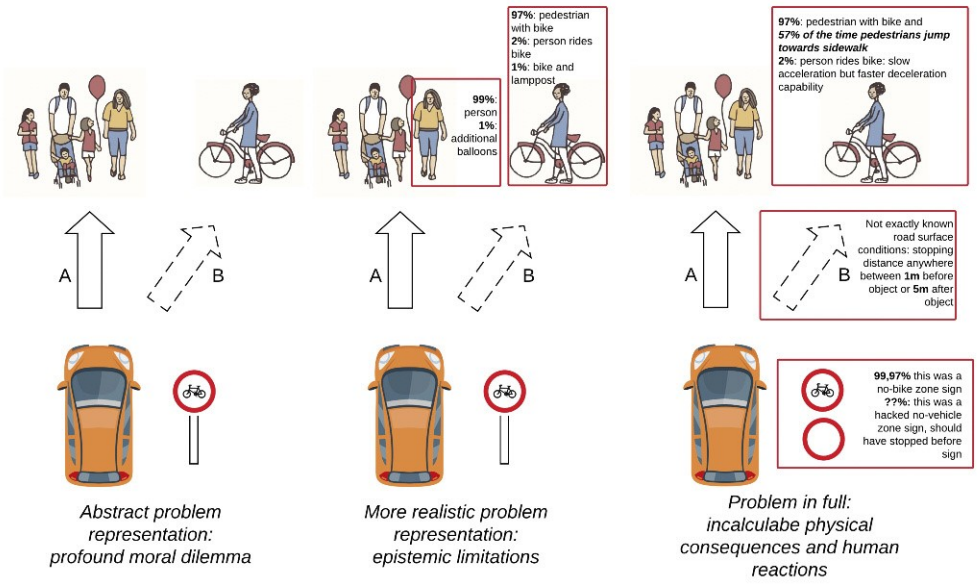

Figure 1: The epistemic constraints of the autonomous car limit the scope of design-time moral investigations.

uncertainty of the object categorisation itself. Also, it has been shown that the neural networks - the technology that performs the identification - can be tricked [2] (the resilience of neural networks to such attacks is a research subject at our department). The uncertainty attached to such situations means that the ethical dilemma itself is only known probabilistically.

At any rate, we are not expecting moral agency from the car itself. Instead these decisions are supposed to be made design-time. Here is where the fallacy of our epistemic facilities come into play. When asked in an experiment a large majority of subjects will say that the vehicle should sacrifice one to save many. But such preventive action has the nontrivial consequence that this known vehicle behaviour allows for malicious actors to trick cars into killing people - by actually jumping in front of a car or even without if the object detection can be tricked. Or, the pedestrian might jump away but the vehicle happens to swerve in the same direction, causing the very tragedy it tried to prevent. When presenting such scenarios to subjects they often backtrack on their previous opinions. Nontrivial consequences are one reason why surveys like the Moral Machine [3] are flawed.

Let us instead entertain a typically contractarian proposal: the autonomous car shall brake intensely in such situations but it will never swerve. This proposal has the marks of good rule-based systems: it is both simple to implement and to understand and results in predictable behaviour.

Such a proposal, as long as we think in the context of the current traffic conditions, would result in tragic casualties in some individual cases, which might have arguably been prevented by a human driver. However, the simplicity of such a self-preserving rule will allow those very conditions to be changed so that the situation won't arise.

The contractarian approach is rational because it does not attempt to solve moral value dilemmas that have proven to be intractable over the last couple of hundred years. It also accounts for the unimaginability of future situations that is the reality of design-time work. What it does instead is come up with a simple set of rules design-time, asking for the consent of all road users, and thereby in run-time it allows for more control of the situations that impact humans by virtue of being easily predictable. This also allows an evolution of the overall attitude of human road users towards autonomous vehicles in yet unforeseen ways to manage their presence in their own self-interest.

Finally, in order for the contractarian approach to work it needs to stick to its principles - beyond simplicity and intelligibility, those behaviour patterns should be well-known or even advertised; it should be accepted if not with full consensus but at least with compromise; and these behaviour patterns should be guaranteed to operate consistently as much as possible. About a hundred years ago, when the automobile was a novelty, pedestrians needed to vacate some parts of the streets in ways they were not required to in the age of horse carriages - but in return they got traffic lights. At a red light, drivers stop even if there is absolutely no traffic for kilometres: the contract is binding and ensures safety by not allowing any self-judged overruling.

This work was supported by the Bolyai scholarship of the Hungarian Academy of Sciences and by the ÚNKP-18-4 New National Excellence Program of the Ministry of Human Capacities.

References:

[1] P. Lin: "Why Ethics Matters for Autonomous Cars", M. Maurer et al. (Hrsg.), Autonomes Fahren, DOI 10.1007/978-3-662-45854-9_4, 2015.
[2] A. Nguyen, J. Yosinski, J. Clune: "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", in Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.
[3] E. Awad, et al.: "The Moral Machine experiment", Nature, 563 (7729), 59, 2018.

Please contact:

Mihály Héder,
mihaly.heder@sztaki.mta.hu
https://www.sztaki.hu/en/science/departments/hbit