Automated Vehicles & Discrimination

Report – February 2020

Executive summary

1. This paper summarises the discussions at a workshop held on the 27 November 2019 organised by the Assuring Autonomy International Programme on behalf of the Centre for Connected and Autonomous Vehicles. It was held at the University of York. The workshop brought together 20 attendees from across academia, industry, and Government. Key terms are defined in Annex A.

2. The purpose of the workshop was to explore to what extent automated vehicles (AVs) should be programmed to detect, and make decisions based upon, characteristics protected by anti-discrimination legislation. The Equality Act 2010 was used to define these protected characteristics.

3. The workshop concluded that presently the sorts of sensory discernments required for the AV to detect, understand, and make decisions based upon, the protected characteristics are not yet technically feasible. Overall system safety and accuracy, as well as a reasonable degree of transparency from industry alongside public engagement, are the primary immediate concerns. In the meantime, AVs should be developed to raise everyone's safety outcomes.

4. The workshop was therefore of the view that it was better to rely on accurate trajectory prediction to avoid impacting objects, and other general safety mechanisms, than identifying protected characteristics to inform decision-making, and AV behaviour should be tested across a diverse range of scenarios.

Discrimination as an ethical concern

5. Discrimination is often addressed under ethics. For the purposes of the workshop ethics was defined as "'standards of behaviour' for society to help avoid harm". The workshop focussed solely on vehicle behaviour and not wider concerns for data privacy or the reasonable adjustments required for service provision.

6. Attendees found it helpful to conceptualise the topic as a discussion of 'input' ethics and 'outcome' ethics. **Input ethics** concerns the decisions made by curators about how the vehicle will be programmed and the process for doing this, such as the training data used. **Output ethics** focuses on the expectations we have for the behaviour of the vehicle, such as allowing all pedestrians time to step out of the road after crossing.

Provisional findings

7. On the whole, it is unclear how to dictate input ethics. Doing so could even be counter-productive to the aims of equality and fairness, considering that:

a. The technology is not yet developed enough for us to predict how it will work in the future,

b. The proxies that are always used to identify protected characteristics (e.g. a wheelchair) could worsen the problem as accuracy would not be 100%,
c. We lack the evidence that controlling input ethics is necessary to identify protected characteristics,
d. It would be difficult to maintain technology neutrality and may stifle innovation,
e. There is a competitive advantage for companies that possess diverse training data sets, and so government intervention may not be needed provided safety outcomes are met.

8. Considering the concerns with setting input ethics, establishing the right outcomes for AV behaviour was considered the best way to approach the problem. This is because:

a. It is the best way to balance safety against innovation,
b. Expecting a reasonable level of transparency both of industry and the systems they build addresses many concerns we have (though how this is done remains unclear),
c. Verification may be the best opportunity to check for ethical concerns, such as whether wheelchair users are recognised as pedestrians.

9. Accuracy may be the best measure for the right outcomes since reliability may not be pertinent when applied to emergent behaviour.

10. Though it won't help to dictate input, industry will need to provide clear reasoning for the choices they have made as part of the input stage.

11. In the event that something goes wrong, it is important that we know why. This includes when things go wrong at an aggregate level. It will be crucial to know whether harm is being caused because of an issue with the programming or for extraneous variables (beyond the control of the curator).

12. Interestingly, precisely because it is assumed that AVs will need to be safety assured and therefore potential discrimination can be tested for, AVs may avoid the issues with 'unembedded AI' discrimination (e.g. worse credit scores for women) especially because the resulting harm is likely to be a lot more serious (and visible) with a vehicle.

13. AVs should raise the standards of safety for everyone not target specific people for better safety outcomes.

14. Questions of whether harm can occur to an individual through discrimination that either does not affect safety outcomes or leads to a better safety outcome are best left to the Courts.

15. It is unclear how the programming of AVs in relation to mapping and location (and the potential for discrimination that arises therein) can be done well. The example was explored where a vehicle either avoids driving through, or avoids stopping to pick up a passenger in, a neighbourhood with high vandalism rates, where this neighbourhood also has a high ethnic minority population. Whether or not this decision, which may have unfair outcomes for ethnic minorities, is objectively justified would ultimately be decided by the Courts.

16. Engaging with the public is crucial to the success of articulating the right outcomes. Public engagement requires the following aspects:

a. Honesty; AVs will not completely eliminate road deaths,
b. Credible voices; who are the people with the evidence?
c. Transparency; what decisions have been made to improve safety?
d. Deliberative/consultative approach; how can the public get involved and shape Government policy?
e. Exposure to the technology; so that people can learn more about AVs in general

17. It will be important to test these findings as the policy area develops and vehicles are deployed without safety drivers.

18. In the future, following the deployment of automated vehicles, when more data is available and vehicles are more sophisticated, it may be possible to achieve better more nuanced safety outcomes through weighting certain amounts of training data or programming the vehicles in a certain way (i.e. addressing input ethics). At this point, on-going comprehensive stakeholder engagement may yield helpful and unexpected insights. However, it is difficult to say if this will be desirable prior to the availability of this data and this level of technological sophistication.

Recommendations

19. Government and industry should engage the public in an honest, transparent, and deliberative manner to help build understanding of the ethical trade-offs involved for automated vehicles and get their input to policy.

20. Government should fund trials of AVs that demonstrate their social benefits to help the public understand how AVs could be a part of their life.

21. Government could support trials of AVs in bus lanes, which are more easily controlled than normal lanes and allows the public to be safely exposed to AV technology.

22. Government should investigate the provision of data for verification and validation of AV behaviour that is suitably representative of individuals with protected characteristics.

23. When programming an AV, industry should consider whether the training data used is representative of a modern and diverse United Kingdom and appropriately detailed. It is important to determine and articulate the purpose of the training data sets, as well as to ensure they are context-sensitive.

24. Industry should rigorously verify vehicle behaviour to ensure that the vehicle properly categorises all instances of human and that it does not discriminate against anyone. With unsupervised learning, training data should be as detailed as possible - it might be that the characteristics that are truly relevant for fair outcomes are much more comprehensive than we currently know.

Centre for Connected & Autonomous Vehicles

ANNEX A – Definitions

- Accuracy: the extent to which the result of a measurement or calculation conforms to the correct value.

- Curator: the individual(s) making decisions about how the automated driving system (ADS) behaves, with an emphasis on the data inputted.

- Emergent behaviour: behaviour of a system which is not apparent from the properties of its constituent parts.

- Reliability: the probability that some action is performed as specified.

- Trajectory prediction: estimating the speed and direction of movement of some object, e.g. a pedestrian or road vehicle.