The Second Workshop on Implementing Machine Ethics

- **Title**: First steps towards an Ethical Agent for checking decision and behaviour of an Autonomous Vehicle on the Rules of the Road
- **Authors**: Gleifer Vaz Alves[1], Louise Dennis[2], Michael Fisher[2]
- **Affiliation**: [1]UTFPR (*Universidade Tecnologica Federal do Parana*), Brazil; [2]University of Liverpool, United Kingdom.

Abstract

Autonomous systems are widely used in different scenarios, e.g. robotic systems, autonomous aircraft, underwater vehicles, autonomous cars. And intelligent agents can be properly used to model the high-level decisions on such autonomous systems [5]. In many cases, these intelligent systems shall act in accordance to some set of rules or guidelines issued by some authority such as a Certification Agency, these are intended to ensure the systems will operate in a "right" way. But, it is not enough to know the system will act "correctly", it is also necessary to assure that even in complex and dynamic scenarios the system will be able to handle ethical dilemmas that may take place. That is the sort of problem that is tackled in [1]. The authors have presented a formal verification of ethical decision-making within autonomous systems controlling autonomous aircraft. The autonomous system is modelled as an agent that has the so-called Rules of the Air programmed in. The agent is supposed to select the most ethical plan (available) according to a set of ethical requirements that are designed to preserve human life, no matter what the Rules of the Air may represent.

The deployment of Autonomous Vehicles (AV) brings several concerns related to regulations, safety, security, liability, etc. With this in mind, the German government has established a national ethics committee for AV's. This committee has composed a document with the German Ethics Code for the deployment of AV's [4]. The document presents 20 Ethical Guidelines (**EG**). Here we (partial) quote some of the guidelines which share some correspondence with our proposal. **EG 1**: *says to improve safety for all road users*; **EG 2**: *mentions that protection of individuals takes precedence overall, it also tells the need to reduce the level of harm*; **EG 5**: AV's *should prevent accidents wherever this is practically possible*; **EG 7**: *in hazardous situations that prove to be unavoidable, protection of human life enjoys top priority*; **EG 8**: *genuine dilemmatic situations (human life decision), where the behaviour is unpredictable, can not be programmed such that they are ethically unquestionable. Technological systems are designed to avoid accidents ... but they can not replace or anticipate the decisions of a responsible driver with the moral capacity to make correct judgements. Yet in this guideline, it is mentioned that it would be desirable to have for regulation an independent public sector agency.* **EG 9**: *in unavoidable accidents ... any (biased) distinction ... is strictly prohibited.* **EG 19**: *in emergencies, the vehicle must autonomously enter into a "safe condition" (handover routines), i.e., when it is necessary to leave the autonomous mode.* Moreover, this Ethics Code [4] provides as an example, a "trolley case", *i.e.* an ethical dilemma where there is no good choice to be selected. In the example, there is a human driver faced with a split-second decision between hitting children playing by the roadside and driving over a cliff. The driver might choose to sacrifice herself. That would be a personal, intuitive decision, and potentially represent a quite complex process of deliberation. Such decision should not be taken by a programmer. We could add to this discussion that in this very same scenario (dilemma), the same person may take different split-second decisions at different points in her life.

It is strongly recommended that AV's should be deployed on the roads only using an approach which can be safe, responsible, ethical and regulated. We have previously observed that the design of an AV usually does not take into account the Rules of the Road (see [7] and [8] for further discussion). That is the reason why we have been studying the UK Highway code

(specifically the Road Junction rules) [3] in order to demonstrate how the Rules of the Road can be formalised and represented for an AV modelled by an agent. Consequently, we should be able to verify the behaviour of such an agent according to the Road Junction rules. As the following step, we are now concerned with the ethical behaviour of this agent, when submitted to an urban traffic environment while making use of Road Junction rules: there are clearly many situations when claiming that you were acting in accordance with the UK Highway Code would not be sufficient answer to an accusation that your actions were unethical.

On previously works we have started to use intelligent agents to model and formally verify the behaviour of an AV. Firstly, simulating an urban-like environment, where the AV-agent is supposed to select reliable decisions considering different levels of emergency that may occur in the simulated environment (see [5] and [6]). Besides, we have formalised a subset of rules from the UK Highway code (the Road Junction rules) using LTL (*Linear Temporal Logic*) [7]. With this, the AV-agent could be implemented according to the UK Rules of the Road [3]. The next step was taken to define an agent-based architecture [8] capable to represent: a Road Junction environment with traffic lights, stop signs, road users (i.e. pedestrian, cyclist, driver, etc), among other elements; the Rules of the Road (formalised in LTL); the AV-agent implemented using a BDI (*Belief-Desire-Intention*) language Gwendolen [2]; and the formal verification of related properties, for instance: *Is it always the case the car will stop at a red traffic light?*

Here, we intend to extend the results previously described in order to support for an *Ethical agent* and also consider the existence of a Human-in-the-loop (HITL). We take as our starting point the work in [1], but consider the Rules of the Road, instead of the Rules of the Air.

It is expected that our Ethical agent would be able to verify the decisions and behaviour of an AV-agent in a Road Junction environment. Specifically, our proposed architecture includes two agents: AV-agent and Ethical agent, which interact within an urban traffic environment through sensors and actuators. With sensors the agent(s) should perceive the urban traffic environment, *e.g.*, sensing a red traffic light and with actuators the agent(s) will act in the environment, *e.g.*, stopping the car at a red traffic light. The Ethical agent is responsible to sense the environment and send suggestions to the AV-agent which may reflect on its decision and behaviour. Our architecture also explicitly includes the HITL since the Ethical agent, in the case of a complex scenario or one representing a dilemma which only a human may decide, must warn the HITL and proceed with a handover routine.

Inspired by Asimov's Laws of Robotics [10][1] we propose the following Laws:

1. The AV-agent will not take an action which it is believed will may cause harm to some Road User.
2. The AV-agent should strictly follow the Road Junction rules unless this conflicts with the First Law.
3. An AV-agent should obey the orders given by the Ethical agent except when such orders would conflict with the First and Second Laws.
4. An Ethical agent must warn the HILT when no action is available that does not contravene one of the first three laws.

To illustrate our proposal we briefly present two scenarios. Scenario #1: in a road junction environment the green traffic light is on, there is a car (embedded with an AV-agent) on the road and it should cross the junction to not block the traffic flow. Nonetheless, there is a distracted road user (*e.g.* a pedestrian crossing the crosswalk). As a result, the AV-agent decides to carefully hit the brakes and stops to not cause any harm to the road user. In this scenario, the AV-agent has followed the suggestion given by the Ethical agent (*by Law* **3**), where it tells the AV-agent to use Law **2** (*i.e.*, it has not followed the Road Junction rules because to do so risked harm to a Road User).

Scenario #2: Imagine the situation from Scenario #1. But, now we should also consider that behind the AV-agent there is a queue of cars quite close to the AV-agent and driving at almost the maximum speed allowed on the road. The Ethical agent would firstly send a suggestion to the AV-agent to hit the brakes, but this would cause potential harm to the cars behind (or even itself) (*by Law* **2**). The second attempt would be send a suggestion to not hit the brakes, however this may cause harm to the pedestrian (also *by Law* **2**). The Ethical agent verifies

[1] Asimov's laws have been critiqued as a basis for machine ethics [9, 11], however we are not using them here to provide an ethical theory but to control the priorities of the HITL, the Ethical Agent and the AV-agent.

the intended behaviour of the AV-agent and checks that in both situations the AV-agent will eventually break the Laws. In this situation, the Ethical agent has no alternative order to suggest to the AV-agent, but since no action is available that obeys all three laws then, by the Law **4** the Ethical Agent orders the AV-agent to give back the control to the human.

As a final remark, we may say that our proposed Laws try to comply with the UK Highway Code and the German Ethical Guidelines. For that, we stress the following: **i.** the Laws consider the role of a HITL since the trolley cases should not be programmed to replace decisions of a responsible driver in split-second scenarios; **ii.** the protection of human life is indeed a top priority; **iii.** also the reduction of the level of harm; **iv.** and the safety for all road users should be taken into account.

Several open issues could be considered in our proposal, for example: **i.** Should we consider Regulations for the Ethical Agent? **ii.** Who would be responsible for such Regulations? **iii.** Could it be an independent public sector agency (as mentioned in [4])? **iv.** Or should it be a sort of Vehicle Certification Authority? **v.** Should the programmed software embedded in the Ethical Agent work like a "black box" (e.g., as in [12])? If so, who would have access to it? **vi.** Could this software be constantly updated with daily data use of the car and the urban traffic environment? Or the software should have all scenarios and behaviours programmed upfront?

References

[1] Dennis, Louise, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems 77* (March 1, 2016): 1-14. https://doi.org/10.1016/j.robot.2015.11.012.

[2] Dennis, L. A. Gwendolen semantics: 2017. Technical Report ULCS-17-001, University of Liverpool, Department of Computer Science, 2017.

[3] Department for Transport. Using the road (159 to 203). The Highway Code - Guidance - GOV.UK, 2017.

[4] Luetge, Christoph. The German Ethics Code for Automated and Connected Driving. Philosophy & Technology 30, no. 4 (December 1, 2017): 547-58. https://doi.org/10.1007/s13347-017-0284-0.

[5] Fernandes, L. E. R.; Custodio, V.; Alves, G. V.; Fisher, M. A Rational Agent Controlling an Autonomous Vehicle: Implementation and Formal Verification. In *Proceedings First Workshop on Formal Verification of Autonomous Vehicles*, Turin, Italy, 19th September 2017, edited by Lukas Bulwahn, Maryam Kamali, and Sven Linker, 257:35-42. **Electronic Proceedings in Theoretical Computer Science**. Open Publishing Association, 2017. https://doi.org/10.4204/EPTCS.257.5.

[6] Alves, G. V.; Dennis, L.; Fernandes, L.; and Fisher, M. Reliable Decision-Making in Autonomous Vehicles. In Leitner, A.; Watzenig, D.; and Ibanez-Guzman, J., editor(s), *Validation and Verification of Automated Systems: Results of the ENABLE-S3 Project*, pages 105-117. Springer International Publishing, Cham, 2020.

[7] Alves, G. V.; Dennis, L.; and Fisher, M. Formalisation of the Rules of the Road for embedding into an Autonomous Vehicle Agent. In International Workshop on Verification and Validation of Autonomous Systems, pages 1-2, Oxford, UK, July 2018.

[8] Alves, G.; Dennis, L.; and Fisher, M. An Agent-Based Architecture supported by Temporal Logic for representing the Rules of the Road on Autonomous Vehicles. In Luckcuck, M.; Farrell, M.; and Fisher, M., editor(s), *FMAS Workshop Pre-Proceedings 2019*, volume 1, pages 41-48, Porto, Portugal, October 2019. **FMAS - A satellite workshop of Formal Methods 2019**, Porto, Portugal.

[9] Anderson, M. and Anderson, S. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28(4):15–26, 2007.

[10] Asimov, I. Runaround. In *Astounding Science Fiction*. Street & Smith, March 1942.

[11] Murphy, R. and Woods, D. Beyond asimov: the three laws of responsible robotics. *IEEE Intelligent Systems*, 24(4), 2009.

[12] Winfield, A.F.T and Jirotka, M. The case for an ethical black box. In *Towards Autonomous Robotic Systems*. Springer, 2017.