An Agent-based architecture with support to Ethical Decisions on a Road Traffic Scenario

Gleifer Vaz Alves[1], Louise Dennis[2,3] and Michael Fisher[2,3]

***Abstract*— The road traffic scenarios with autonomous vehicles pose many challenges when automating the behaviour of Autonomous Vehicles (AVs) to obey road traffic rules. Is it possible to assure that a given AV will always respect the road traffic rules? More challenging, is it always the case that the vehicle is supposed to actually respect such rules? How about those scenarios where a driver needs to break some rule to avoid potential harm to a road user? That is the reason we need to further study the behaviour of autonomous driving in a road traffic environment. With this in mind, we use intelligent agents to model the behaviour of an AV and add an additional ethical agent. These agents belong to an architecture, where a HITL (Human In The Loop) is also present. The HITL must be ready to retake the vehicle control in complex scenarios. In this paper, we introduce these elements and also describe two simple examples. In further work, we shall implement our architecture using a BDI agent programming language that gives support to the formal verification of desired properties.**

I. INTRODUCTION

Autonomous systems are widely used in different scenarios, e.g. robotic systems, autonomous aircraft, underwater vehicles, Autonomous Vehicles (AVs). And intelligent agents can be properly used to model the high-level decisions on such autonomous systems [1]. In many cases, these intelligent systems shall act in accordance with some set of rules, principles or guidelines issued by some authority such as a Certification Agency. These are intended to ensure the systems will operate in the "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 issues. Indeed as noted by Nallur [2], human society will likely only trust autonomous systems if it is known to have a set of moral (ethical) principles that guide and constrain its behaviour (cf. [3]).

That is the sort of problem that is tackled in [4]. 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.

[1]Federal University of Technology—Parana (UTFPR), Ponta Grossa 84017-220, PR, Brazil gleifer@utfpr.edu.br

[2]Department of Computer Science, University of Manchester, Manchester M13 9PL, UK {louise.dennis, michael.fisher}@manchester.ac.uk

[3] Dennis and Fisher's work was funded by EPSRC Grant EP/V026801/1 *Trustworthy Autonomous Systems Verifiability Node*.

The design of Autonomous Vehicles (AVs) includes several stages like sensing the environment, communication, object detection, and many concerns related to regulations, security, liability, safety, etc. The safety of an AV is related to different aspects such as collision avoidance, road user safety, transparency (considering decisions and actions taken by the AV), the proper use of road traffic rules, etc. Notice that transparency of autonomous systems is a required feature that enables levels of certification to the operation of autonomous systems [5]. Therefore, to assure the safety behaviour of an AV, the system should have (among other capabilities) transparency, meanwhile, transparency is also one of the key issues to determine the ethical behaviour of an autonomous system [5].

As outlined by Prakken [6] and Alves et al. [1], the design of AVs should include the proper deployment of the road traffic rules. In [7], an agent-based architecture is presented, where a subset of road junction rules is modelled, implemented, and formally verified using model checking techniques. With this, an AV can have its behaviour verified against road traffic rules and we can shed a light on the question: *Can an AV equipped with traffic rules drive safely on the roads?* But, as previously mentioned, we still need to guarantee the AV will abide by not only the traffic rules but also some ethical guidelines to obtain the necessary level of trust for the stakeholders. Therefore, we aim to add an extra layer (an ethical one) to the previously mentioned architecture, so that the AV behaviour on the roads, when using the traffic rules can be verified considering a set of ethical principles. Specifically, we will create an ethical agent to work along the AV-agent already existent in the architecture deployed by Alves et al. [7].

Moor [8] defines four types of artificial moral agents: agents with ethical impact, implicit ethical, explicit ethical and full ethical agents. In [9], a clear distinction between implicit and explicit agents is described, where the former has been programmed to behave ethically without an explicit representation of ethical principles, while the latter uses ethical principles and operates on the basis of this knowledge. When building an ethical system there are different approaches, e.g. top-down and bottom-up (cf. [10]). Recently, in [11] the authors have outlined two additional approaches: i. a governor module style ethical system and ii. a global ethical system, in the former ethics is placed in a sub-system or constrains the actions of the rest of the system, while in the latter the ethical reasoning is involved in all system reasoning, where all decisions are considered ethical.

In our work, we choose to use an explicit agent and a

governor module style ethical system, since the ethical agent has its ethical guidelines explicitly defined and the agent is the only component of our architecture responsible for the ethical reasoning and decisions but uses its reasoning to constrain the actions of the rest of the system.

An ethical system usually has a moral theory to guide its ethical decisions. There are different ethical theories for agents such as virtue, deontological and consequentialist theories [10]. Considering the latter, perhaps the most used type is the utilitarianism theory, where the morally right action is the one that follows the rule that is deemed to produce the most favourable balance between good over bad. By using utilitarianism we can define utilities to examine how much good an action does. In our approach, we shall use an ethical agent based on utilitarianism since such theory seems a suitable moral theory to abstract the consequences related to the application of traffic rules. Specifically, we define a set of utility functions to capture different situations that may take place in a given road traffic scenario. The utility functions results can be compared so that the ethical agent selects the maximise utility.

In this work, we intend to extend the results previously described in [7] in a way that support for ethical decisions is properly addressed. Besides, the ethical agent, we also consider the presence of a Human-in-the-loop (HITL). We take as our starting point the work in [4], but consider the Rules of the Road, instead of the Rules of the Air. It is expected that our ethical agent can be able to verify the decisions and behaviour of an AV-agent on a road traffic environment. The architecture proposed here can be seen as an extension to the architecture defined by Alves et al. [7], where there are two agents: AV-agent and ethical agent, which interact within an urban traffic environment through sensors and actuators. With sensors the agents should perceive the environment, *e.g.*, sensing a red traffic light and with actuators, the agents will act in the environment, *e.g.*, stopping the car at a red traffic light. The ethical agent is responsible for sensing the environment and send recommendations to the AV-agent (based on the ethical principles programmed into the ethical agent), which may reflect on the decisions and behaviour of the AV-agent. Our work explicitly includes the HITL since the ethical agent, in the case of a complex scenario or an ethical dilemma, where it is advisable that only a human may decide what to do, must warn the HITL and proceed with a handover routine.

The remainder of this paper is organised as follows. Section II presents the ethical agent definition and the corresponding proposed architecture, including the set of utilities that shall be used by the ethical agent. Next, Section III describes the communication between the agents and the ethical reasoning. Section IV illustrates two simple examples. And Section V points out our final remarks.

II. ETHICAL AGENT

In this section, we present the details of our ethical agent proposal including the architecture, principle, utilities and definition.

Fig. 1 shows the so-called Ethical Agent Architecture (**EAA**) which presents the following elements:

- Urban Traffic Environment: includes the agents and artifacts (*e.g.*, stop sign, pedestrian, traffic light).
- AV-agent: models the AVs behaviour, which can obtain percepts with its the sensors and act in the environment via actuators (*e.g.*, hit the brakes, turn left, stop the car). NB: we consider the AV-agent is based on the one implemented in [7].
- Ethical agent: responsible for monitoring the environment and the AV-agent. With this, this agent does the ethical reasoning and send recommendations or warnings to the AV-agent or the HITL.
- Human in the Loop (HITL): models the human driver which can communicate with the AV-agent and the ethical agent.

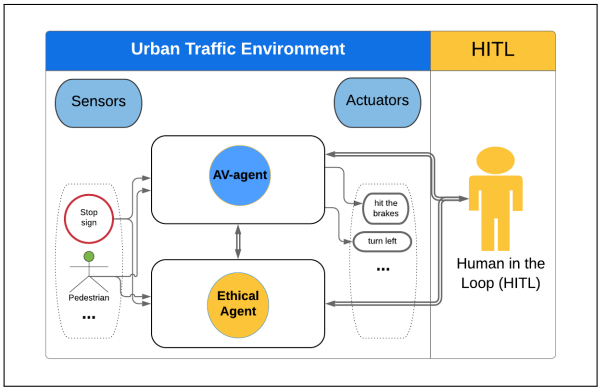

Fig. 1. Ethical Agent Architecture (EAA)

Notice we choose to split up the kind of reasoning into two agents: AV-agent and ethical agent. So that the ethical reasoning is executed by a specific agent and not mixed up with the agent responsible to model the AV behaviour. With this modular approach, we intend at some point to extend the **EAA** architecture to support multiple AV-agents.

A. Principles and Utilities

As mentioned earlier, the deployment of AVs is related to regulations, safety, security, liability, etc. That is why, the German government has established a national ethics committee for AVs. This committee has composed a document with the German Ethics Code for the deployment of AVs [12]. The document presents 20 Ethical Guidelines (**EG**). Here we (partial) quote some of the guidelines which are used as references to determine the utilities functions for our ethical agent. **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**: AVs *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*.

Notice that **EG 1**, **2**, **5**, and **7** are used as a basis to determine the utilities functions, while **EG 8**, **9**, and **19** are related to the EAA (previously seen in Fig. 1).

Definition 2.1 (Utilities): An ethical agent has a set (U_{tl}) of n ethical utilities:

$$U_{tl} = \{u_1, u_2, \ldots, u_n\}$$

where $n \geq 0$ and each u_i is an element from U_{tl} determined by the application of a single utility function, as follows:

1) `ru_safe`$= 10$ (*the road users are safe*).
2) `av_pass_safe`$= 10$ (*the passengers in the AV are safe*).
3) `av_respects_rule`$= 9$ (*the AV-agent has respected the traffic rules*).
4) `av_damage`$= -1$ (*there is a minor damage in the AV*).
5) `traffic_env_damage`$= -2$ (*there is a minor damage in some object existent the traffic environment, e.g. a traffic sign was broken*).

NB: a given U_{tl} set is established to a certain road traffic scenario when a new scenario is generated, a new U_{tl} set is also obtained, and the previous utility values are dismissed.

B. Definition

As it follows, we give the definition for the ethical agent.

Definition 2.2 (Ethical Agent): An ethical agent (**EA**) is given by the tuple:

$$\textbf{EA} := (U_{te}, B_e, R_r, P_l, A_{ct}, U_{tl})$$

where,

- U_{te} is the Urban Traffic Environment where the agents are placed; the environment has artifacts related to road traffic scenarios, as for example: stop sign, pedestrian, traffic light, cross walk, etc.
- B_e is a set of Beliefs from the agent; there are two kinds of beliefs: **Ethical Beliefs** and **Road Traffic Beliefs**. An example of the former is `noHarmRoadUser`, *i.e.* the agent believes there is no possible harm to the road user, whereas an example of the latter is `greenLightOn`, *i.e.* the agent believes the green light (from the traffic light) is on.
- R_r stands for a set of Road Rules which are programmed into the agent to give the necessary knowledge on the road traffic rules.
- P_l is a set of Plans that are triggered every time the ethical agent is supposed to act and send recommendations.
- A_{ct} is a set of Actions available to the agent.
- U_{tl} is a set of Utilities functions that are calculated by the agent according to Def. 2.1.

NB1: for each plan belonging to P_l, there is a corresponding u_i element that belongs to the U_{tl} set. So that, the ethical agent will select the plan according to the respective values from the ethical utilities. E.g., a plan p_1 has $u_1 = 10$, while a plan p_2 has $u_2 = 9$, so $u_1 > u_2$ and p_1 will be selected by the ethical agent as a recommendation.

NB2: when the obtained ethical utilities are the same, i.e. $u_1 = u_2$, there is uncertainty. As a consequence, a handover manoeuvre to the HITL is engaged, cf. subsection III-B and example 1 (at section IV).

III. ETHICAL REASONING

A. Communication

According to the **EAA** the ethical agent communicates to the AV-agent and the HITL. We define the corresponding Ethical Communication Protocol (**ECP**) which in general is used to transmit recommendations on ethical issues.

Definition 3.1 (Ethical Communication Protocol): the **ECP** is given by the following tuple:

$$\textbf{ECP} := (Sender, Receiver, Env_{pos}, A_{ct}, U_{tl})$$

where,

- Sender: the agent who is sending the message.
- Receiver: the agent who is receiving the message.
- Env_{pos} defines the position in the Urban Traffic Environment related to the actions included the message, e.g. at 1st road junction.
- A_{ct}: the set of actions which is being recommended by the Sender.
- U_{tl}: the set of utilities which has been used by the Sender.

B. Reasoning

The agents presented in the Ethical Architecture (see Fig. 1) can communicate each other, while the ethical agent is responsible for performing some sort of ethical reasoning, these elements are illustrated in Figures 2 and 3, which respectively describe the reasoning for the ethical agent and AV-agent.

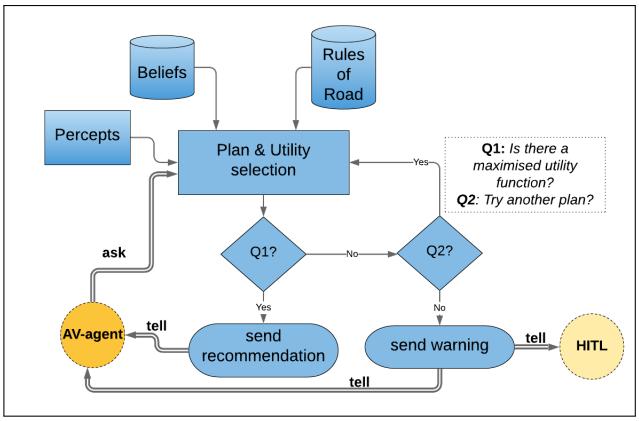

Fig. 2. Ethical agent reasoning and communication

The reasoning flow for the ethical agent can be summarised as follows:

1) The ethical agent selects plans and utilities according to its set of Beliefs, environment percepts and the embedded Rules of the Road.
2) When a given plan and utility are selected it will trigger some recommendation.
3) Next, there is a first query (**Q1**), where it should be checked by comparing two utilities functions, if there is a maximised function.
   a) If so, then a recommendation will be send to the AV-agent.
   b) Otherwise, a second query (**Q2**) is executed to check whether the ethical agent should try another plan or send a warning warning message to the HITL.
   NB: this warning message is a recommendation and should also be sent to the AV-agent, because it will trigger a handover manoeuvre from the AV-agent to the HITL.

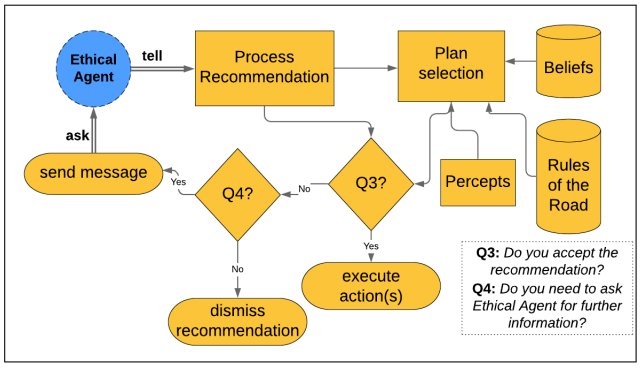

Fig. 3. AV-agent reasoning and communication

The reasoning flow for the AV-agent can be summarised as follows:

1) The AV-agent receives some recommendation from the ethical agent and should process the received message, checking each field according to the message structure (see Def 3.1).
2) Next, there is a query (**Q3**) for the AV-agent, where it should decide whether or not to accept the received recommendation.
3) If AV-agent chooses to accept. The suggested actions are performed and the reasoning flow is done.
4) Otherwise, an additional query (**Q4**) is prompted where the AV-agent may request further information for the ethical agent.
   a) If AV-agent decides to ask for more information, then a message is sent to the ethical agent.
   And the agent will eventually returns to step **1** as soon as it receives a new message from the ethical agent.
   Example: AV-agent may ask ethical agent using the following message: `ECP(AV,ET,RJ-1,?,?)` symbol ? means the Sender is asking to the Receiver for a different set of actions and utilities, respectively. This is possible because the ethical agent selects a plan according to the utility functions. So, probably the ethical agent has a set of candidate plans. And the AV-agent agent may ask for a given action and utility function that has not been selected by the ethical agent. This can be helpful for the reasoning process of the AV-agent. Perhaps, the AV-agent may select another plan that would override the initial recommendation sent by the ethical agent. The AV-agent may also request additional information, because the recommendation (sent by the ethical agent) may lead the AV-agent to a conflict situation, for example, in this case the plan override can be a suitable solution.
   b) Otherwise, the AV-agent chooses to completely dismiss the recommendation sent by the ethical agent and the reasoning flow is ended.
5) Moreover, the AV-agent has the plan selection stage which is connected to the beliefs, the percepts (received from the environment) and the set of road traffic rules (the rules of the road).

The plan selection stage is directly connected to the recommendation process, since a new recommendation may trigger a different plan for the agent. Besides, the trigger of plan has consequences in query (**Q3**), where a given knowledge may lead the agent to accept or not a recommendation.

NB: the AV-agent previously implemented in [7] has only the stages mentioned on item **5** (above). Thus, we will need to extend this agent with a communication capability (with the ethical agent) and a recommendation process stage.

IV. EXAMPLE

Here, we describe two simple examples to show how our system is supposed to work according to the reasoning and communication mechanisms.

In these examples, we assume the existence of the following elements:

- Two agents: AV-agent and ethical agent (`ET`), which are placed on an Urban Traffic Environment.
- The HITL.
- The scenarios present a road junction with a car (the AV-agent) and a pedestrian (or a road user).
- Also the agents are programmed with the set of road traffic rules.
- **Example 1**:

```
<ET perceives the Green Light at
RJ-1; Pedestrian in a hurry crossing;
Car behind>
<ET selects two plans and
the corresponding utility functions>
<u_i = 10>; <u_j = 10>
<ET sends a recommendation to HITL>
ECP(ET, HITL, RJ-1, {handover manouervre,
```

```
retake control}, {u_i, u_j})
```

In the first example both utility functions (u_i and u_j) have the same value (10), which correspond to plans where the road user should be safe and also the AV passenger. Thus, the ethical agent faces a dilemma and should warn the HITL.

- **Example 2**:

```
<ET perceives the Green Light at
RJ-2; Pedestrian in a hurry crossing>
<ET selects two plans and
the corresponding utility functions>
<u_i = 10>; <u_j = 9>
<ET sends a recommendation to AV>
ECP(ET, AV, RJ-2, {hit the brakes},
{u_i})
```

The second example illustrated a similar situation (considering **example 1**). However, at road junction 2 (**RJ-2**) there is only a pedestrian in a hurry, there is no car behind the AV-agent. So, the selected utility functions (u_i and u_j) have different values (resp. 10 and 9), which correspond to the utility where the road user should be safe and to the case where the AV-agent has respected the traffic rule. Since, $u_i \geq u_j$, then u_i is selected and the ethical agent sends a recommendation to the AV, which is to *hit the brakes* to assure the road user safety. Notice that by accepting this recommendation, the AV-agent will not cross the junction with the green light, i.e. it will not respect a traffic rule.

V. CONCLUSION

In this paper, we have presented an agent-based architecture, which has an ethical agent as the main element. This agent is explicit and uses a utilitarianism approach to determine utility functions that help to establish ethical guidelines. These guidelines are mainly designed to guarantee the safety of road users. The ethical agent is responsible for monitoring the road junction environment and an AV-agent. According to the obtained perceptions, the ethical agent uses its reasoning process to trigger recommendations that are sent to the AV-agent and/or to the HITL.

Our following step will be the implementation of the Ethical Agent Architecture, where we intend to connect the ethical agent module into the agent-based architecture previously developed by Alves et al. [7]. For that, we will use GWENDOLEN agent programming language [13] and MCAPL framework [14]. With this, we will be able to specify formal properties using linear temporal logic and run the model checker AJPF to verify these properties. We aim to specify properties that represent the behaviour and decisions taken by the agents. For example, *is it always the case that when the ethical agent sends a given recommendation r_1, then r_1 leads to scenarios where a given road user is safe?* Apart from that, we will also enable that our ethical agent monitors the decisions taken by the AV-agent, which correspond to the use of road traffic rules, since the AV-agent implemented in [7] makes use of such rules.

Moreover, we aim to extend our **EAA** architecture (presented in Fig. 1) to a multi-agent approach, where an ethical agent can monitor two or more AV-agents. That is why the **EAA** architecture splits the sort of reasoning executed by the agents, i.e. the ethical reasoning and the autonomous driving functions based on the road traffic rules. Notice that a multi-agent architecture will bring some challenges to overcome, such as how to solve possible conflicts that arise between different recommendations sent from the ethical agent to different AV-agents in the same road junction environment.

REFERENCES

[1] G. V. Alves, L. Dennis, and M. Fisher, "Formalisation and Implementation of Road Junction Rules on an Autonomous Vehicle Modelled as an Agent," in *Formal Methods. FM 2019 International Workshops*, ser. Lecture Notes in Computer Science, E. Sekerinski, N. Moreira, J. N. Oliveira, D. Ratiu, R. Guidotti, M. Farrell, M. Luckcuck, D. Marmsoler, J. Campos, T. Astarte, L. Gonnord, A. Cerone, L. Couto, B. Dongol, M. Kutrib, P. Monteiro, and D. Delmas, Eds. Cham: Springer International Publishing, 2020, pp. 217–232.

[2] V. Nallur, "Landscape of Machine Implemented Ethics," *Science and Engineering Ethics*, vol. 26, no. 5, pp. 2381–2399, Oct. 2020, arXiv: 2009.00335. [Online]. Available: http://arxiv.org/abs/2009.00335

[3] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, June 2016, publisher: American Association for the Advancement of Science Section: Reports. [Online]. Available: https://science.sciencemag.org/content/352/6293/1573

[4] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889015003000

[5] A. F. T. Winfield, S. Booth, L. A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R. I. Muttram, J. I. Olszewska, F. Rajabiyazdi, A. Theodorou, M. A. Underwood, R. H. Wortham, and E. Watson, "IEEE P7001: A Proposed Standard on Transparency," *Frontiers in Robotics and AI*, vol. 0, 2021, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2021.665729/full

[6] H. Prakken, "On the problem of making autonomous vehicles conform to traffic law," *Artificial Intelligence and Law*, vol. 25, no. 3, pp. 341–363, Sept. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s10506-017-9210-0

[7] G. V. Alves, L. Dennis, and M. Fisher, "A double-level model checking approach for an agent-based autonomous vehicle and road junction regulations," *Journal of Sensor and Actuator Networks*, vol. 10, no. 3, 2021. [Online]. Available: https://www.mdpi.com/2224-2708/10/3/41

[8] J. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, 2006.

[9] M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine*, vol. 28, no. 4, pp. 15–15, Dec. 2007, number: 4. [Online]. Available: https://www.aaai.org/ojs/index.php/aimagazine/article/view/2065

[10] L. A. Dennis and M. Slavkovik, "Machines That Know Right And Cannot Do Wrong: The Theory and Practice of Machine Ethics," p. 4, 2018.

[11] V. Nallur, L. Dennis, S. Bringsjord, and N. Govindarajulu, "A Partially Synthesized Position on the Automation of Machine Ethics," *Journal of Philosophy and Technology*, vol. to appear, 2021.

[12] C. Luetge, "The German Ethics Code for Automated and Connected Driving," *Philosophy & Technology*, vol. 30, no. 4, pp. 547–558, Dec. 2017. [Online]. Available: https://doi.org/10.1007/s13347-017-0284-0

[13] L. A. Dennis, "Gwendolen semantics: 2017," University of Liverpool, Department of Computer Science, Tech. Rep. ULCS-17-001, 2017.

[14] L. A. Dennis, M. Fisher, M. P. Webster, and R. H. Bordini, "Model Checking Agent Programming Languages," *Automated Software Engineering*, vol. 19, no. 1, pp. 5–63, Mar. 2012. [Online]. Available: https://doi.org/10.1007/s10515-011-0088-x