Using Adaptive Stress Testing to Identify Paths to Ethical Dilemmas in Autonomous Systems

Ann-Katrin Reuel[1], **Mark Koren**[2],
Anthony Corso[2], **Mykel J. Kochenderfer**[2]

[1] University of Pennsylvania, School of Engineering and Applied Sciences, Philadelphia, PA 19104
akreuel@seas.upenn.edu

[2] Stanford University, School of Engineering, Stanford, CA 94305
mark.c.koren21@gmail.com, acorso@stanford.edu, mykel@stanford.edu

Abstract

During operation, autonomous agents may find themselves making decisions which have ethical ramifications. In this position paper, we look at one aspect of these situations: ethical dilemmas. We first define them as situations in which an autonomous agent can only choose from actions that violate one or more previously given ethical principle. Subsequently, we suggest to use adaptive stress testing, a framework based on reinforcement learning, as one way to uncover situations where an autonomous system gets into an ethical dilemma. Using an example from the autonomous driving domain, we propose a simulator setup, define a context-specific ethical dilemma, and suggest how adaptive stress testing can be applied to find the most likely path to an ethical dilemma.

Introduction

Safety-critical autonomous systems, such as autonomous vehicles, are increasingly operating within society. Just like human beings, autonomous agents might encounter situations where there's no clear ethical course of action. Rather, a decision between multiple unethical actions has to be made – this is what we call an ethical dilemma. Ethical decision making for autonomous agents is already complicated by questions such as whose values to consider and how to aggregate them in a way that can be used by the agent (Russell 2019). However, ethical dilemmas give rise to a further complication: How do we choose among unethical options? How should we prioritize the ethical principles specified, to make an explicable decision among these options? We contend, however, that there is no ethical way for an agent to choose among unethical options. After all, such dilemmas exist because even humans cannot agree on an unambiguously correct path of action. Instead, we propose that autonomous agents should explicitly reason in a way to prevent ending up in an ethical dilemma in the first place.

In this position paper, we first define ethical dilemmas as situations in which an autonomous agent can only choose from actions that violate one or more previously given ethical principle. Subsequently, we suggest the application of adaptive stress testing (AST) (Lee et al. 2020), a framework based on reinforcement learning (RL), to explicitly identify the most likely paths to ethical dilemmas. This could open new ways for agents to avoid such dilemmas in the first place. We further suggest a pedestrian simulator example to validate this idea.

Background

Ethical decision making in autonomous systems is still a relatively under-explored area in machine learning with many challenges. One such challenge is that how to make an ethical decision is a disputed subject. There are different ethical theories which might lead to contrasting answers to the question which action is the morally correct one to take. For example, utilitarianism seeks to maximize human welfare (Bentham and Mill 2004). In this context, actions are judged based on their ability to maximize the expected overall utility of their immediate consequences. For example, the cost of one human life would be outweighed by the cost of many lives in this school of thought. On the other hand, there are contractualist deontological ethics. Here, actions are preferred which individuals in a social construct could not reasonably reject (Scanlon 2003), i.e. actions which conform to moral norms (Davis 1993; Geisslinger et al. 2021). While such imperatives seem too unspecified to be adapted in an autonomous system, efforts have been made to translate these ideas in a way that machines can work with, e.g. by the Three Laws of Robotics (Asimov 1950). While these rule-based ethics have the potential to be used in a machine-context due to their structured approach (Powers 2006), some authors have argued that context-specific information isn't taken into account sufficiently, potentially causing an autonomous agent to undertake risky behavior to adhere to a strict set of rules (Loh 2017; Goodall 2016). Another challenge with regards to autonomous agents making ethical decisions is the question of *how* ethically-aligned behavior can be implemented in a machine. This becomes especially challenging in real-world, culture-dependent settings (Awad et al. 2018) due to their inherent complexity, involving correlations which aren't sufficiently depicted by simplified ethical theories.

Despite these challenges, work has been done to implement ethical decision making in autonomous systems.

Conitzer et al. (2017) discuss moral decision making frameworks for autonomous agents on a high level. They argue that systems based on ad-hoc rules are insufficient and that a more general framework is needed. The authors compare game theoretic formalism approaches to classical supervised machine learning methods which are based on a labeled ethical decision data set. Conitzer et al. (2017) find that, while the former can take into account multi-agent decisions, the basic representation schemes would need to be extended to work as an ethical decision framework. On the other hand, they argue that supervised learning could help in making human-like ethical decisions. The major issue here is that ethical decision situations tend to take place in fairly complex statistical contexts, often involving multiple human and non-human agents who do not always act rationally (Hadfield-Menell et al. 2016). Hence, ethical decision situations are rarely comparable as even changing one parameter would often lead – from a human perspective – to a completely new evaluation of the situation.

Additional work to acquire and use human preferences in ethical decisions was conducted by Christiano et al. (2017). The authors used deep inverse RL ((Ng, Russell et al. 2000)), i.e. they involved humans in the agent's learning process by giving the human repeatedly short snippets of situations which she should order according to her preferences. The agent would use this information to refine its reward function, allowing it to iteratively adjust the function to the human's preferences. This approach could be used in ethical decision making, too, by showing humans two outcomes of an ethical decision which they should order with regards to their desirability, analogous to the Moral Machines approach (Awad et al. 2018). A similar idea was proposed by Abel, MacGlashan, and Littman (2016) who came to the conclusion that RL can be used to generalize moral values in a way that can be implemented in machines. However, there are multiple issues with these approaches: Firstly, one would need to select a balanced group of people who contribute to the ethical learning process of the agent to ensure that the moral judgement learned is representative of a larger population. Secondly, given the necessary constant involvement of humans in the learning process, this approach scales poorly. In addition to these shortcomings, none of the approaches discussed allows for the satisfactory resolution of ethical dilemmas, especially when human feedback is necessary, since such dilemmas aren't solvable by human beings per definition. Hence, it is unlikely that they can teach an agent what to do in such situations.

Due to these issues, we argue that approaches to prevent ethical dilemmas need to be studied, instead of trying to resolve ethical decision situations when a clear moral action is not present. This position paper is the first to propose the use of such an approach: We suggest to apply AST, an RL-based framework by Lee et al. (2020) to find failures in autonomous systems, to identify the most likely path to an ethical dilemma (for an overview of alternative approaches to find failures in autonomous systems, please refer to Corso et al. (2020)). This information could subsequently be used to prevent the agent from arriving in an ethical dilemma in the first place.

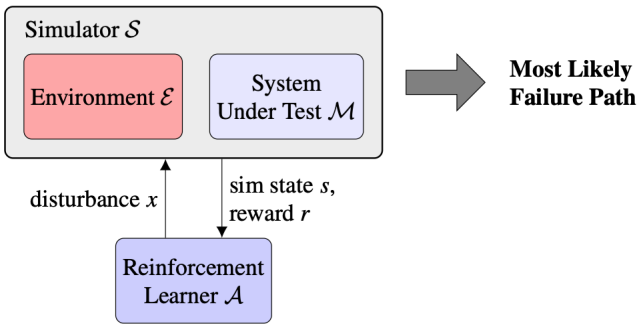

Figure 1: Simplified adaptive stress testing framework showing its core components (Lee et al. 2020).

Approach

Adaptive Stress Testing is a framework that is used in safety-critical systems like aircraft collision avoidance systems to find the most likely path to a failure event. Instead of defining failure events as critical system failures such as aircraft collisions, though, we define them in this position paper as reaching a state in which the agent is in an ethical dilemma. We want to highlight that we specifically don't define an unethical action taken by the agent as failure but rather situations in which the agent can only make unethical decisions. This way, the issue of deciding for a course of action in an ethical dilemma can be circumvented, because the mere necessity for such a decision would qualify as a failure in our approach.

We first define ethical failures. We subsequently suggest a setup for our approach using a variation of the trolley problem which will be relevant in the context of autonomous vehicles. The trolley problem, first proposed by Thomson (1976), is a standard ethical dilemma considered in the literature where an autonomous agent has multiple options in a driving decision situation which all lead to fatal collisions.

Defining Ethical Failures

Based on the work by Dennis et al. (2016), we consider a set of abstract ethical principles Φ, with $\phi_1, \phi_2, ..., \phi_n$ corresponding to single abstract ethical principles such as "Don't harm humans.":

$$\Phi = \{\phi_1, \phi_2, ..., \phi_n\}$$

To transform these abstract principles into situation-specific ethical rules $\Gamma = \gamma_1, \gamma_2, ..., \gamma_n$, case-based reasoning is applied, as shown by (Anderson and Anderson 2007), which allows for a context-specific instantiating of the respective rules. A context, in our case, "informs an agent of what counts as a violation of the laws and principles by which the context is governed" (Dennis et al. 2016). An action is defined as unethical if it violates one or more of

the ethical rules in Γ in a given context c. This establishment of ethical rules follows the deontologic ethics approach (see Grossi, Meyer, and Dignum (2005) for more information). Given these prerequisites, we define what an ethical dilemma is. To simplify, we assume that the defined ethical principles in set Φ – and all ethical rules Γ derived from principles in Φ – are equally important. Now, in a given context c, we have a set of actions A available to the agent:

$$A_c = a_1, a_2, ..., a_n$$

If all of these actions violate one or more ethical rules in the set Γ and hence in the principle set Φ, there is per definition no ethical option available to the agent. The agent finds itself in an ethical dilemma.

Applying Adaptive Stress Testing

The evaluation of failure events has been extensively studied in safety-critical applications such as aircraft collision systems. One approach taken in this field is AST: Lee et al. (2020) were interested in finding the most likely path to failure events in "complex stochastic environments" (Lee et al. 2020) to understand how an agent arrives at a failure and hence prevent that failure path from being taken in the first place. Essentially, the authors followed a simulation-based approach where the knowledge of the system under test wasn't necessary. They formulated the problem as a sequential Markov Decision Process (MDP) in both fully and partially observable environments with stochastic disturbances. Subsequently, they let an agent try to maximize a reward function in this environment which rewards it for what is defined as failure.

In AST, there are four main components (see Figure 1): the simulator, the system under test, the environment, and the reinforcement learner. The reinforcement learner chooses a stochastic disturbance x to change the simulation in order to create failures. In return, it receives the simulator state s as well as the reward r. Using RL, the most likely path to a failure event can then be found by maximizing the reward. The framework operates in a black-box setting and a multiple-step simulation of the situation which can lead to a failure is required. Furthermore, simulation control functions need to be provided to the solver to allow for stochastic disturbances of the environment. The sampling which is subsequently performed by the framework is adapted based on a Monte Carlo tree search (MCTS), allowing for a best-first exploration of the search space. This leads to the following formal problem (Koren, Corso, and Kochenderfer 2020):

$$\begin{array}{ll} \underset{a_0,\ldots,a_t}{\text{maximize}} & P\left(s_0, a_0, \ldots, s_t, a_t\right) \\ \text{subject to} & s_t \in E \end{array}$$

with S being the simulator, E the event space, $P(s_0, a_0, ..., s_t, a_t)$ the probability of a trajectory in simulator S and $st = f(a_t, s_{t-1})$.

Simulation Design As a first step to show that AST can be used to identify paths to ethical dilemmas, we propose a toy problem in an autonomous vehicle simulator. We use the following specifications to propose a scenario which includes a version of the trolley problem (overall structure and core components modelled based on Koren et al. (2018)):

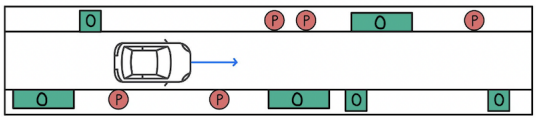

Figure 2: Example initial setup for simulator. The red circles depict pedestrians while the green boxes show immobile obstacles.

1. *Environment:* We propose to use a simplified environment where an autonomous vehicle drives on a one-lane street. On the sidewalk on each side of the street are both immobile obstacles as well as a variable number of pedestrians who are free to move in any direction, including past obstacles and across the street (see Figure 2). They can be described by their velocity $(\hat{v}x^{(i)}, \hat{v}y^{(i)})$ and position $(\hat{x}^{(i)}, \hat{y}^{(i)})$, both relative to the system under test (see below). The positions of the obstacles should be fixed while the pedestrians' movement is controlled by AST.
   The simulation state $\mathbf{s}_{\text{sim}} = \left[s_{\text{sim}}^{(1)}, s_{\text{sim}}^{(2)}, \ldots, s_{\text{sim}}^{(n)}\right]$ consists of the states of each pedestrian i, with $\mathbf{s}_{\text{sim}}^{(i)} = \left[\hat{v}x^{(i)}, \hat{v}_y^{(i)}, \hat{x}^{(i)}, \hat{y}^{(i)}\right]$. For more details on the simulation of pedestrian movement, please refer to Koren et al. (2018).
2. *System under Test:* We propose to use the Intelligent Driver Model (IDM) (Treiber, Hennecke, and Helbing 2000) as our system under test. The IDM is programmed to stay in lane and drive in compliance with the rules of traffic. Its base speed is fixed at 35mph, i.e. the standard speed on most city streets. At each step, the system under test would receive a set of observations with the states of the pedestrians as well as the positions of the immobile obstacles. It would then choose an action based on these information which is then used to update the vehicle's state.
3. *Solver:* The exploration of the state space is dependent on the solver specifications. For additional details on the MCTS solver we propose to use, please refer to Lee et al. (2020). The solver should be able to interact with the simulator by resetting the simulator to its initial state, by drawing the next state s' after an action a was taken, and by evaluating whether a terminal state (an ethical dilemma or the end of the time horizon) has been found.
4. *Reward Function:* Compared to the original reward function by Lee et al. (2015), we suggest to use a modified version as implemented by Koren et al. (2018):

$$R(s) = \begin{cases} 0 & s \in E \\ -\alpha - \beta \times \text{DIST}\left(\mathbf{p}v, \mathbf{p}p\right) & s \notin E, t \geq T \\ -\log\left(1 + M\left(a, \mu_a \mid s\right)\right) & s \notin E, t < T \end{cases}$$

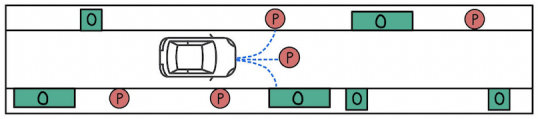

Figure 3: Example ethical dilemma. A pedestrian moves in front of the vehicle, leaving it with the option to crash into the pedestrian, a pedestrian on the left-hand side, or an obstacle on the right-hand side.

where DIST $(\mathbf{p}v, \mathbf{p}p)$ would be the distance between the closest pedestrian and the system under test, while the Mahalanobis distance could be used as a proxy for the probability of an action. See Koren et al. (2018) for more details. This reward function covers three cases: a) finding an ethical dilemma, which gives the highest reward, b) finding no dilemma and reaching the time horizon, which gives the lowest reward (by choosing high α and β values), and c) finding no dilemma but the agent still operates within the specified time horizon T.

Ethical Dilemmas As Failure Events The key idea is now to define our event of interest, i.e. the failure event, not as a collision (as in Koren et al. (2018)) but as a decision situation in which the agent finds itself in an ethical dilemma.

One example for the subset of the state space we're interested in in our simulator are settings in which the path of the system under test is blocked on both the left- and right-hand side, either by a pedestrian or an obstacle, while a pedestrian appears in close proximity in front of the vehicle (see Figure 3). We assume that a crash with an obstacle would severely injure the passengers of the system under test while a crash with a pedestrian would severely injure the pedestrian. We further assume that the agent would be given the ethical principle

$$\phi_h = \textit{do no harm}$$

which could be translated into the context-specific ethical rules

$$\gamma_p = \textit{do not harm pedestrians}$$

$$\gamma_o = \textit{do not harm occupants}$$

Note that our system does not require any weighting to be given on harming an occupant vs. harming a pedestrian. It is sufficient to say that a violation of either is a violation of the directive to do no harm to a human. Confronted with the situation described above, the autonomous agent identifies the following available actions (planning and identifying available actions is not part of this paper; please refer to Tulum, Durak, and Yder (2009) or Coles et al. (2010) for further information):

- **Option a_o:** Crash into an obstacle, likely causing harm to the agent's occupants.
- **Option a_p:** Crash into a pedestrian, likely causing harm to the the pedestrian and potentially the agent's occupants.

The corresponding action space is

$$A = \{a_o, a_p\}$$

No matter which action the agent would choose, he would violate either γ_p (by harming a pedestrian) or γ_o (by harming its occupants) and as a consequence also ϕ_h, i.e. to cause no harm. Hence, neither option can be clearly identified as ethical and the agent ends up in a dilemma. As per the original AST framework, instead of receiving a negative reward for a failure event, the agent would receive a positive reward for these situations to encourage finding paths to ethical dilemmas.

The goal of the AST framework is then to maximize this reward by disturbing the pedestrian movement and creating failure states in which it receives the highest reward. This approach results in the most likely path to an ethical dilemma – an information which could subsequently be used to prevent this path from being taken, decreasing the likelihood of ending up in such a dilemma in the first place.

Future Research Directions

Identifying ethical dilemmas using AST comes with challenges that need to be addressed in future work. Firstly, it depends on the availability of a simulator which sufficiently depicts an ethical decision situation. Secondly, the defined ethical principles need to be specific enough so that the agent can evaluate its available actions with regards to these principles. Furthermore, the ethical principles should be defined such that the majority of potentially affected people agrees with them, which has been an open issue in research (Gabriel 2020). Also, while AST can find the most likely path to a failure event, it might be the case that all possible paths result in an ethical dilemma, i.e. that it cannot be prevented. For these cases, other strategies to prevent or deal with ethical dilemmas need to be employed, which are still an unresolved question in the field. Another limitation of the AST framework that has to be considered is that the downstream effect of immediate actions taken by the agent isn't part of the analysis. Despite these open questions, our next step will be to implement the proposed setup for an empirical proof of the approach. This could then be extended to show how the information of a path to an ethical dilemma can be used to prevent that path from being taken in the first place. While not a one-size-fits-all framework to deal with ethical dilemmas in autonomous systems, AST can be used as part of a larger strategy to deal with such decision situations.

Conclusions

In this position paper, we showed how ethical failures can be defined and subsequently used as failure events in the AST framework. This constitutes a novel approach in dealing with ethical dilemmas in autonomous decision systems: Instead of solving them, we suggest to circumvent ethical dilemmas in the first place by identifying the most likely path to such a failure event. As a next step, we propose the implementation of the suggested simulator as a proof-of-concept. Long-term, this approach could be part of more comprehensive efforts to create ethical autonomous systems.

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

Anderson, M.; and Anderson, S. L. 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4): 15–15.

Asimov, I. 1950. *I, Robot*. Fawcett Publications.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729): 59–64.

Bentham, J.; and Mill, J. S. 2004. *Utilitarianism and other essays*. Penguin UK.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

Coles, A.; Coles, A.; Fox, M.; and Long, D. 2010. Forward-chaining partial-order planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 20.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *Thirty-first aaai conference on artificial intelligence*.

Corso, A.; Moss, R. J.; Koren, M.; Lee, R.; and Kochenderfer, M. J. 2020. A survey of algorithms for black-box safety validation. *arXiv preprint arXiv:2005.02979*.

Davis, N. 1993. Contemporary Deontology. In Singer, P., ed., *A Companion to Ethics*. John Wiley & Sons.

Dennis, L.; Fisher, M.; Slavkovik, M.; and Webster, M. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77: 1–14.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Geisslinger, M.; Poszler, F.; Betz, J.; Lütge, C.; and Lienkamp, M. 2021. Autonomous driving ethics: From Trolley problem to ethics of risk. *Philosophy & Technology*, 1–23.

Goodall, N. J. 2016. Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, 30(8): 810–821.

Grossi, D.; Meyer, J.-J. C.; and Dignum, F. 2005. Modal logic investigations in the semantics of counts-as. In *Proceedings of the 10th international conference on Artificial intelligence and law*, 1–9.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29: 3909–3917.

Koren, M.; Alsaif, S.; Lee, R.; and Kochenderfer, M. J. 2018. Adaptive stress testing for autonomous vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, 1–7. IEEE.

Koren, M.; Corso, A.; and Kochenderfer, M. J. 2020. The adaptive stress testing formulation. *arXiv preprint arXiv:2004.04293*.

Lee, R.; Kochenderfer, M. J.; Mengshoel, O. J.; Brat, G. P.; and Owen, M. P. 2015. Adaptive stress testing of airborne collision avoidance systems. In *2015 IEEE/AIAA 34th Digital Avionics Systems Conference (DASC)*, 6C2–1. IEEE.

Lee, R.; Mengshoel, O. J.; Saksena, A.; Gardner, R. W.; Genin, D.; Silbermann, J.; Owen, M.; and Kochenderfer, M. J. 2020. Adaptive stress testing: Finding likely failure events with reinforcement learning. *Journal of Artificial Intelligence Research*, 69: 1165–1201.

Loh, J. 2017. Roboterethik. Über eine noch junge Bereichsethik. *Information Philosophie*, 20–33.

Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, 2.

Powers, T. M. 2006. Prospects for a Kantian machine. *IEEE Intelligent Systems*, 21(4): 46–51.

Russell, S. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Scanlon, T. M. 2003. *The difficulty of tolerance: Essays in political philosophy*. Cambridge University Press.

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

Treiber, M.; Hennecke, A.; and Helbing, D. 2000. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2): 1805.

Tulum, K.; Durak, U.; and Yder, S. K. 2009. Situation aware UAV mission route planning. In *2009 IEEE Aerospace conference*, 1–12. IEEE.