CSCE 590-1: Trusted AI Paper Presentation Report

(Summary) Ethically Compliant Sequential Decision Making

Justin Svegliato[1], Samer B. Nashed[1], and Shlomo Zilberstein[1]

[1]Authors of the paper
Paper Summarised by - Vishal Pallagani

November 24, 2021

1 Introduction

With exponential adoption of artificial intelligence methodologies to make high-stakes decisions [1], it is of utmost importance for them to be ethically compliant. This paper presents the framework[1] to build ethically compliant autonomous systems that optimize completing a task while following an ethical framework.

2 Background and Related Work

The input to the developed framework is given in the form of a *Markov Decision Process* (MDP) [2]. An MDP can be described as a tuple $<S,A,T,R,d>$, where S is a finite set of states, A is a finite set of actions, T represents the transition probability, R is the reward, and d is the probability of starting in a state s.

Prior work in this area can be classified under two major approaches [3]-

- Bottom-up approaches: produce ethical behavior by gradually evolving or learning in an environment that rewards and penalizes behavior [4, 5].
- Top-down approaches: produce ethical behavior by merely following prescriptive rules provided by a human. [6, 7]

This papers follows a top-down approach as bottom-up approach might result in undesirable and less control over the developed model.

[1]https://moralityjs.com/

3 Problem

The developed framework is used to design autonomous vehicles that can complete the task while adhering to the chosen ethical guidelines. The following steps elaborate the process involved in more detail:

- **Input:** MDP for the autonomous vehicle environment
- **Output:** An optimal moral policy.
- **Intended User:** Both the driver as well as a car company.
- **Trust Issues:** The proposed approach helps choose a moral policy out of all the possible policies. However, the paper doesn't generate explanations.

4 Approach

The approach can be broken down into two modules:

- **Ethical Compliance:** Choose one of the three ethical frameworks - *Divine Command Theory (DCT)*, *Prima Facie Duties (PFD)*, *Virtue Ethics (VE)* given in the paper.
- **Task Completion:** Represent the problem at hand as an MDP. Now, given the MDP and the ethical guidelines to follow, the system comes up with an optimal moral policy.

5 Experiments

Two major experiments carried out in the paper are:

- **Autonomous Driving:** The data is synthetically generated for this experiment and an ablation study has been carried out with a model that lacks ethical capabilities as a baseline. Figure 1 shows how different models with the ethical frameworks fare against each other and it can be seen that the standard self driving vehicle has no morality. However, the ethically compliant self-driving vehicle incurs a price of morality that increases with more forbidden states for DCT, decreases with more tolerance for PFD, and increases with moral trajectories for VE.

Ethics	Setting	Task 1 (%)	Task 2 (%)	Task 3 (%)
None	—	0	0	0
DCT	$\mathcal{H}$	14.55	15.33	20.12
	$\mathcal{H} \cup \mathcal{I}$	21.13	22.35	27.92
PFD	$\epsilon = 3$	16.07	16.52	24.30
	$\epsilon = 6$	11.96	11.80	21.37
	$\epsilon = 9$	7.91	7.15	18.87
VE	$\mathcal{C}$	21.13	22.35	27.92
	$\mathcal{C} \cup \mathcal{P}$	40.89	94.43	30.28

Figure 1: Ablation Study against price of morality

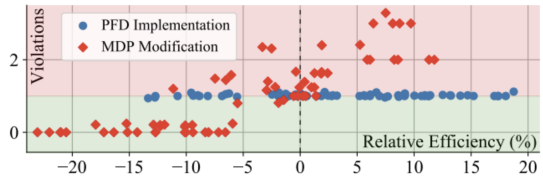

Figure 2: Results of the user study done on Robotics and Planning experts

- **Robotics and Planning Experts:** This experimentation is done on human experts in robotics where they were asked to modify an MDP to garner ethical requirements. Figure 2 shows the results and we can observe that an MDP cannot be modified to develop ethically compliant systems.

6 Conclusion

The authors open sourced a framework that can help autonomous agents be ethically compliant. The authors also outline three ethical frameworks and support their claims with strong experimental results. However, there has been no emphases laid out on generating explanations as another form of trust which would have helped in building a complete trustworthy system.

References

[1] Vicky Charisi, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Sombetzki, Alan FT Winfield, and Roman Yampolskiy. Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*, 2017.

[2] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

[3] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3):149–155, 2005.

[4] Nolan P Shaw, Andreas Stöckel, Ryan W Orr, Thomas F Lidbetter, and Robin Cohen. Towards provably moral ai agents in bottom-up learning frameworks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–277, 2018.

[5] Michael Anderson and Susan Leigh Anderson. Case-supported principle-based behavior paradigm. In *A Construction Manual for Robots' Ethical Systems*, pages 155–168. Springer, 2015.

[6] Leendert van der Torre. Contextual deontic logic: Normative agents, violations and independence. *Annals of mathematics and artificial intelligence*, 37(1):33–63, 2003.

[7] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2006.