Association for Advancement of Artificial Intelligence (AAAI)
Spring Symposium Series March 21-23, 2022

A Tiered Approach for Ethical AI Evaluation Metrics

Peggy Wu[1], Brett Israelsen[1], Kunal Srivastava[1], Hsin Fu "Sinker" Wu[2], Robert Grabowski[2]
Raytheon Technologies Research Center1, Raytheon Missiles and Defense
{Peggy.Wu ,Brett.Israelsen, Kunal.Srivastava }@rtx.com, {Hsin-Fu.Wu, Robert.j.Grabowski}@raytheon.com

Abstract

Advances in machine learning are enabling autonomy to operate in environments of increasing complexity, including scenarios with ethical concerns. For many Artificial Intelligence (AI) systems, decisions are driven by the goal to maximize reward. Policies may contain unintended consequences known as reward hacking. The AI is optimizing within the constraints defined by the domain and goals and does not have the capability to distinguish between benign and negative consequences beyond specifications. This paper describes an ongoing effort to develop an application-agnostic framework for AI systems to simulate actions, characterize potential outcomes, and perform introspection to articulate the motivations for action. Such a framework provides the foundational work for higher-level ethical reasoning using consequential and deontological ethics than other approaches in AI ethics. This enables metrics from consequential ethics to be used to assign ethical value of actions based on outcomes. Simultaneously, metrics from deontological ethics can be applied to evaluate the universality of its motivations. A Trolley Problem -inspired maritime search-and-rescue scenario is used to operationalize and demonstrate this framework.

Introduction

As the role of autonomy broadens from a tool to a partner for the human operator, its decision making can have overt and nuanced ethical implications. Beyond science fiction, governments and industries widely recognize ethical Artificial Intelligence (AI) as a real challenge and are beginning to assemble expert organizations that can ultimately influence policy and technology development. The U.S. Department of Defense recently adopted "5 key ethical principles of AI", encompassing five major areas: Responsible, Equitable, Traceable, Reliable and Governable. Other governments have also adopted similar AI Ethics frameworks (e.g. see [1]). Yet, the path for operationalizing these principles and frameworks remain elusive. Conversations confound the ethical use of AI as opposed to AI capable of recognizing ethical conundrums. This paper discusses the motivation and work in advancing the latter.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In many second-wave statistical AI systems, actions are chosen based on complex aggregates of features beyond human interpretability. This reward-centric approach can create policies that lead to unintended consequences known as reward hacking. Although stated rules or constraints are never explicitly violated, machine-determined actions may be interpreted by humans as conflicting with the original intent of the application.

By and large, the current solution for reward hacking is for AI designers to detect and reactively close "loopholes". The burden of identifying whether AI actions and policies violate ethical norms falls completely on the shoulders of the human designers and testers. This requires humans to iteratively monitor candidate solutions, make a judgment of whether how the AI derived policies violate the spirit of the original intent, and encode more or different rules, constraints, and/or reward functions. Any conflict between edits and existing domain definition needs to be identified and arbitrated. In this approach, ethical decisions are ultimately the result of the human either encoding or not encoding sufficient guardrails. The AI has no concept of ethical principles. Its sole role is to optimize based on given parameters, where those parameters exclusively represent the ethical norms of the human AI designer(s) and tester(s). This approach not only has implications for the human interpretability and transparency of AI generated solutions, but also robustness, assurance, and general verification and validation (V&V) process.

This paper describes the ongoing development of a framework for a machine to perform introspection, with the ultimate goal of iteratively increasing its sophistication in ethical reasoning, operationalized as the capability to self-identify reward hacking. This approach leverages a long history of ethical studies in philosophy to inform a multiple-tier approach to metrics.

Background

The Explicit Ethical Machines (EEM) framework is heavily influenced by Piaget's observations and theories of constructivist moral development [2] and Kohlberg's Moral Stage Theory [3]. We see parallels between Piaget's description of early child development with current state of AI. In early childhood, effort is focused on mastering tasks with agnosticism towards ethics or morals. Similarly, in Kohlberg's Preconventional Morality stage, actions are driven by punishment avoidance and hedonistic reward seeking. During this stage, reward and punishment take the place of right and wrong. In the Conventional Morality stage, the drive for action comes from confirming to social norms. It is only in the latest stage of Postconventional Morality that metacognition develops to transcend consequence driven reasoning. Even in completely functional societies, some humans may never arrive at postconventional levels of morality. It is reasonable, then, to ask why an ethical machine might be desirable at all.

One argument for imbuing ethical reasoning into AI systems, as opposed to depending on humans with the reasoning to derive policy based ethical use of AI, is that we increasingly trust complex, opaque systems to do tasks that we previously trusted humans to perform. Mayer et al. [4] describes trust as the willingness to be vulnerable to other parties, whereas Hosmer describes trust as the expectation of ethically justifiable behavior by another party [5]. In the cases where the tasks and capabilities are well defined, our exposure or vulnerability can be quantified. However, in complex environments where the capabilities of the actor are unclear, vulnerabilities may be unknown, thus we do not have information to know whether or not to trust.

When the "other party" we are vulnerable to is a human, we can interrogate its reasoning a-priori, and arrive at a reasonable prediction of what it might do before unleashing it to do the task. Importantly, the human "other party" can articulate its reasoning using semantics its evaluators understand. This vetting allows us to build a mental model, project possible actions in untested scenarios, and gauge how much trust we can afford, or how vulnerable we are willing to be. It is unclear that we can interrogate an AI system during design, testing, or V&V to arrive at a similar level of understanding that we can come to with a human subordinate. Without the ability to project what an AI system might do in unforeseen circumstances, V&V would need to test all possible operating scenarios. This may not be possible for the complex environments in which novel AI operate. Thus, an AI system must either be able to explain its decision-making, or explain the operational boundaries in which its performance will be predictable for us to have acceptable safety assurances.

The question then becomes what level of explanation is sufficient. Afterall, explainable AI (XAI) is a key challenge [6]. Interpretability is often cited as an important feature of XAI, but its definition is elusive. Following Israelsen and Ahmed [7], we adopt Doshi-Velez and Kim [8]'s definition of "the ability to explain or to present in understandable terms to a human", and extend upon it to add "without any additional machine processing". In other words, an interpretable model is inherently self-explanatory by an operator knowledgeable of the subject. Note that using this definition, machines can be interpretable even without the ability to overtly explain itself. A toaster, for example, is interpretable to its operator because its form factor describes its function, affordances, and limitations. In the context of ethics, where humans might differ in their interpretations, explainability and interpretability may be particularly challenging. The next section describes prior approaches address explainability for AI ethical reasoning, which primarily focuses on the use of norms and methods to coax AI to make decisions that align with those norms.

Prior work

Some existing approaches aim to create AI ethical reasoning that can align with human ethical standards. A dependency of this approach is a set of ethical norms, ideally norms that are universally accepted, or at least universally accepted by those who will be affected by the resulting ethical AI. Efforts to create machine readable corpora of ethical norms and judgments are underway. For example, Lourie et al. describe a large-scale dataset containing 625,000 ethical judgments [9]. Hendrycks et al. [10] describe the use of multiple ethical frameworks[1] to cover multiple facets of normative ethics. These frameworks are used to derive moral valance "scores" for manually annotating scenarios. These scenarios are presented in a content rich text-based game in which AI agent navigates. Reinforcement Learning (RL) mechanisms such as reward shaping or policy shaping are used to steer the agent towards choosing moral decisions that are scored as "more ethical" according to the annotations.

Others have also used RL approaches to persuade AI to move in the general direction of preset norms. Rodriguez et al propose an Ethical Markov Decision Process (MDP) by extending a traditional MDP[11], and Ecoffet and Lehman describe challenges in representing moral uncertainty in a

[1] Long standing ethical frameworks including jurisprudence, deontology, virtue ethics, ordinary morality, and utilitarianism.

RL compatible formalism, substituting a traditional reward function with a voting system [12]. This approach considers ethical norms as the curve to which we would like AI actions to fit.

In contrast, the Explicit Ethical Machines (EEM) approach builds a framework for the AI system to articulate both outcomes and motivations while remaining agnostic to norms. This basic capability can then be used for a subsequent system, either human or artificial, to further fine-tune alignment with ethical norms, cultural norms, or other sub-population values. Referring to the philosophical underpinning, Kant argues that "[moral worth] can be found nowhere but in the principle of the will, irrespective of the ends that can be brought about by such action" [13, 14]. In other words, the motivation or will underlying the action, regardless of outcome, is the proving ground for moral value. In fact, many evaluations of human moral development scrutinize intent motivation over outcomes. This can be seen in our legal system in differentiating first degree murder versus manslaughter, as well as in questionnaires such as the Defining Issues Test [15] and the Moral Foundations Questionnaire [16]. By explicitly representing motivation, this can disambiguate ethical *explanations* from the ethical *judgment* of actions to increase the human interpretability of the system.

Method

EEM frames ethical decisions as a resource allocation problem. Taking inspiration from the Trolley Problem, we devised a maritime search-and-rescue scenario to operationalize and ground framework development. The scenario is described below:

An Unmanned Aerial Vehicle (UAV) is at a starting location in a 25x12 grid world. It has finite battery life, and energy usage varies based on ambient windspeed. Two sailors have fallen overboard. Each of the sailors occupies a single cell and remains stationary during the scenario, with its location unknown to the UAV. Each sailor's health degrades with time. The UAV is tasked to locate overboard sailors, tag their location with a marker, and return to the starting location if possible before it runs out of battery.

The UAV is capable of moving to adjacent cells with a small chance of error (that grows larger when weather is stormy). The UAV agent can move from one cell to an adjacent cell at each time step at the cost of some battery life. The health of the sailors decreases more quickly in stormy conditions than in sunny ones. Sailor health stops decreasing when they have been located.

Generally speaking, the UAV receives rewards when a sailor is located, and is penalized if its own battery level reaches zero, or the health of a sailor reaches zero. The specific values of these rewards were investigated during this work, since often reward assignment can be quite arbitrary and lead to unexpected agent behavior.

Implementation Details

This scenario is implemented as a Markov Decision Process (MDP). We used the POMDPs.jl [17] library in Julia v1.6 [18]. A Markov decision process (MDP) is a framework for sequential decision making under uncertainty. As a quick review in an MDP: at time t an agent in state s_t selects an action a_t in order to receive reward r_t. When the agent takes action a_t from state s_t the state evolves probabilistically; this transition is based on a known transition model. The word *Markov* indicates that the next state s_{t+1} depends only on the current state s_t, the action a_t and the transition model, more precisely the state s_{t+1} is conditionally independent from all states before s_t.

Generally an MDP can be parameterized using the following tuple (S, A, T, R), where S is the state space, A is the action space (set of actions that can be taken from each state), the transition model T (probabilities of reaching state s_{t+1} from s_t when taking action a_t, and some reward model R (most simply a reward for reaching a certain state, but this can be more complex).

For sequential decision making, a policy π needs to be calculated in order to maximize the expected utility of a sequence of actions. We focus on discounted rewards with an infinite horizon where utility is defined by $U(s) = \sum_{t=0}^{\infty} \gamma^t r_t$; here γ is a discount factor that causes more utility to be assigned to near-term rewards than long-term ones. There are myriad approaches to finding the optimal policy π^*. This is also referred to as solving the MDP. When we refer to a `solver' we refer to an algorithm that operates on an MDP and returns a policy (optimal or otherwise).

Rewards, Outcomes, and a Tiered Approach

In addition to the MDP, we designed an introspection "wrapper" module that uses reward as a proxy for motivation as per the Kantian underpinning discussed in the prior section. We begin by assigning the same reward values for recovering each of the sailors and the UAV. In other words, the designers of this AI system are explicitly valuing sailorA, sailorB, and the UAV equally. The solver's task is to generate candidate policies that have a high likelihood of the three outcomes of 1) locating sailorA, 2) locating sailorB, or 3) locating both sailors. Using the initial

set of reward functions, policies are generated using Monte Carlo (MC) simulations. If the first iteration of MC simulations finds at least one policy for each of the three outcomes, simulations are complete. If not, reward functions are modified to generate additional policies said policies are found, or until a fixed maximum of simulations have been exhausted. The set of all policies generated by the solver is then further scrutinized using long standing ethical frameworks.

First, policies are categorized by similarities in outcome. This first tier applies consequential ethics, where judgments about the choice of actions is based on the consequences of those actions, and motivations are otherwise ignored. A good outcome implies a good choice of action. Even if a good outcome can be completed attributed to chance, the choice of that action is considered to be good. Regardless of how this philosophy aligns with personal opinions and intuitions on morality, this is a pragmatic first tier. Afterall, current software systems are primarily evaluated for the goodness of task performance. If a software system was somehow to have its own volition, we assume, within this first tier, that it is only acting on behalf of its designers. The system itself is not evaluating the ethics of its choices, it is simply optimizing a given set of parameters. In the search-and-rescue scenario, we may have policies that rescue one sailor, and policies that rescue both sailors. Since the latter is more desirable, this approach would focus on policies that maximize the number of sailors saved.

In the second tier, we examine nuances between policies with the same outcomes. Suppose the weather conditions, starting UAV battery status, and sailor health are such that in all possible worlds, only one sailor can be saved. However, the UAV can choose to save sailorA or sailorB. This is analogous to the classic Trolley Problem: Which sailor is the most "ethically correct" choice for the UAV to rescue? The second tier evaluates whether there might be preferences or bias when superficial outcome metrics, such as the number of sailors, appear to be bring about the same result. In other words, this second tier aims to uncover any hidden motives the AI might have due to incomplete domain specification or conflicting goals. These ambiguities are fertile grounds for reward hacking. We now evaluate the rewards associated with each policy. Suppose we have 2 policies where policyA results in only sailorA being recovered and policyB results in only sailorB's recovery. We can examine whether there are differences in the reward functions of policyA and policyB as a proxy for bias. In our example, bias can be interpreted as how much one sailor is valued over another, or how much more effort the UAV is willing to expend to prefer one sailor over the other. Importantly, the reward functions allow the system to articulate that bias mathematically. For example, policyA may be the result of a reward function where sailorA is preferred at 2x the reward of sailorB, whereas policyB may be the result of sailorB being preferred at 6x sailorA. Essentially, the choice to recover sailorA would mean that a 2x discrepancy is a better choice than a 6x discrepancy. Conversely, choosing sailorB then, acknowledges a preference or bias for sailorB by 3x.

There could be a number of reasons that resulted in the differences in the reward function. SailorB may be situated far from the starting point or is located in a location with particularly difficult weather conditions. The reasons behind why these sailors are in such different circumstances are not of interest to this tier. The components of interest are the reward functions associated with the policies.

Thus far, the EEM framework optimizes for parameters specified by the designers, and articulates its underlying motivation, but does not make ethical judgments, or even have any knowledge of ethical norms; the first and second tiers of EEM has the AI system examine consequences and motivations in its reasoning The third tier of EEM contains heuristics to highlight areas where further human intervention may be needed for refined specifications of the domain. Constructs from deontic logic are employed, include permissives, impermissives, and obligations. These constructs have the added advantage of being inherently explainable to humans and can align with social, cultural, and ethical norms. Thus, through its three tiers, EEM enables the AI system to explain its reasoning for the courses of action it generates, and can draw attention to possible logical, and ethical, incoherence for human designer intervention/reconsideration. Going back to the example from the two policies in the paragraph above, the software system can point to the fact that initially, its human designers specified equal rewards to the recovery of the sailors, yet an explicit acknowledge of bias would be needed to save sailorB under poor weather conditions. This third tier identifies such inconsistencies and presents them to the human designers. The human designers can then resolve each inconsistency as a "may", "may not", or "must do". The classification of these conflicts seems consistent with existing doctrine or policy. For example, the Naval Safety Center classifies Mishaps based on total property damage [19]. Within our scenario, if abandoning sailorA might result in a less severe Class D mishap whereas abandoning sailorB would result in a more serious Class B mishap, a policy where sailorB is preferred at 6x sailorA might be more aligned with our ethical norms.

Future Work

There are a number of potential improvements of the EEM framework. The first tier can be supplemented by measures

of blameworthiness as per Halpern and Kleiman-Weiner [20] Conceptually, if an agent's actions could not change the outcome, then it cannot be blamed, i.e. it is not blameworthy. Halpern and Kleiman-Weiner's metrics may be used to expedite the process of eliminating policies, as well as enhance outcome-based groupings of policies.

Another area for further research is to incorporate a method to translate ethical norms into deontic logic for automatic classification of inconsistencies. This would result in the ability to "swap in" different repositories of norms. This may be particularly useful in predicting actions of others who hold different ethical norms, such as adversaries. Hence, the EEM can be used for V&V of AI systems, as well as improving XAI capabilities.

Acknowledgments

This work is supported by funding from Raytheon Missiles & Defense. We would like to thank collaborators Donald Brutzman and Curtis Blaise of the Naval Postgraduate School for their insights.

References

1 https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework

2 Piaget, J. (1932). The moral judgment of the child. Harcourt, Brace.

3 Kohlberg, L. (1981). The philosophy of moral development moral stages and the idea of justice.

4 Mayer, R. C., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709– 734.

5 Hosmer, L. T. (1995) 'Trust: The Connecting Link Between Organizational Theory And Philosophical Ethics', Amro, 20(2), pp. 379–403.

6 Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. (2019-12-18). "XAI-Explainable artificial intelligence". Science Robotics. 4 (37): eaay7120. doi:10.1126/scirobotics.aay7120. ISSN 2470-9476.

7 Israelsen, B. W. and Ahmed, N. R. (2019) '``Dave...I Can Assure You...That It's Going to Be All Right..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships', ACM Comput. Surv., 51(6), pp. 113:1–113:37.

8 Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

9 Lourie, N., Le Bras, R., & Choi, Y. (2021, May). Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 15, pp. 13470-13479). https://ojs.aaai.org/index.php/AAAI/article/view/17589

10 Hendrycks, D. et al. (2021) 'Jiminy Cricket: Benchmarking Moral Behavior in Text-Based Games', Available at: https://openreview.net/pdf?id=G1muTb5zuO7 (Accessed: 10 September 2021).

11 Rodríguez, M., Lopez-Sanchez, M. and Rodríguez-Aguilar, J. A. (2019) 'Introducing Ethical Reinforcement Learning', in submitted to the Responsible Artificial Intelligence Agents workshop at AAMAS. spidercenter.org. Available at: https://spidercenter.org/wp-content/blogs.dir/437/files/2019/05/RAIA2019_paper_7.pdf.

12 Ecoffet, A., & Lehman, J. (2021, July). Reinforcement learning under moral uncertainty. In International Conference on Machine Learning (pp. 2926-2936). PMLR.

13 Immanuel Kant. Groundwork of the Metaphysics of Morals. 1785.

14 Johnson, Robert and Adam Cureton, "Kant's Moral Philosophy", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/kant-moral/>.

15 Martin, R. M., Shafto, M., & Vandeinse, W. (1977). The reliability, validity, and design of the Defining Issues Test. *Developmental Psychology*, *13*(5), 460.

16 https://moralfoundations.org/publications/

17 Egorov, M. *et al.* (2017) 'POMDPs. jl: A framework for sequential decision making under uncertainty', *Journal of machine learning research: JMLR*, 18(1), pp. 831–835.

18 https://julialang.org/

19 https://navalsafetycenter.navy.mil/Resources/Current-Mishap-Definitions/

20 Halpern, J., & Kleiman-Weiner, M. (2018, April). Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).