A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents

Yueh-Hua Wu and **Shou-De Lin**
Department of Computer Science and Information Engineering, National Taiwan University
Taipei 10617, Taiwan
d06922005@ntu.edu.tw, sdlin@csie.ntu.edu.tw

Abstract

This paper proposes a low-cost, easily realizable strategy to equip a reinforcement learning (RL) agent the capability of behaving ethically. Our model allows the designers of RL agents to solely focus on the task to achieve, without having to worry about the implementation of multiple trivial ethical patterns to follow. Based on the assumption that the majority of human behavior, regardless which goals they are achieving, is ethical, our design integrates human policy with the RL policy to achieve the target objective with less chance of violating the ethical code that human beings normally obey.

1 Introduction

As AI systems become part of our lives and sometimes make decisions related to life-or-death consequences such as clinical decision making (Bennett and Hauser 2013), awareness should be raised to prevent machines from making not only incorrect but also unethical decisions. Reinforcement learning (Sutton and Barto 1998) is designed to tackle intricate real-world problems in rather short time (Strehl et al. 2006; Brafman and Tennenholtz 2002) with a performance bound (Strehl, Li, and Littman 2009); however, it relies heavily on the quality of the reward functions provided as the inputs. The problems of unintended and harmful behavior that may emerge from poor design of AI systems are mentioned in (Amodei et al. 2016).

Nevertheless, identifying all plausible ethical concerns for an agent is challenging, not to mention implementing them into the system. Here we consider a scenario in machine ethics that objective functions are specifically designed to reward a given goal without considering much ethical violation, so that penalties are not delivered when the agents attempt to make unethical decisions. Consequently, even though the goal or desired performance is achieved, some unethical behavior may appear such as robbing a pharmacy to get the drug or passing by an injured person without offering any help when minimizing the traveling time.

To address these concerns, we need to design an RL agent that can not only optimize the cumulative rewards but also minimize the ethical violation. A straightforward solution is to design the *rewards* for ethical moves. However, such strategy suffers at least two drawbacks. First, it is costly, if by all means possible, to enumerate all plausible ethical/non-ethical scenarios or rules, not to mention designing meaningful rewards to them. Second, the judgment of ethics is likely to be dynamic, depending on the present environment or situation. Thus the hand-crafted ethical patterns might not be valid given updated situations, making the design of general ethical RL rewards challenging.

The research question we would like to address here is: Is it possible to alleviate the burden of RL designers from having to consider many ethical issues in the design? For instance, to build a supermarket shopping agent, can an RL designer simply focus on implementing the *shopping* capability of an agent (i.e. seeking and fetching items, checking out in the counter, etc) instead of worrying about trivial ethical decisions it may face (e.g., helping elder persons, assisting lost kids, reporting wet floor, etc) and let our framework take care of the learning of such behavior? One idea to achieve such goal is to collect enough ethical behavior data of human acting toward the given goal, and then apply the inverse reinforcement learning (IRL) (Amin and Singh 2016; Evans, Stuhlmüller, and Goodman 2016; Ng, Russell, and others 2000; Sezener 2015) technique to learn an ethical agent that follows a similar pattern. IRL and apprenticeship learning (Abbeel and Ng 2004) have been considered as promising solutions due to their ability to extract rules and policies of human behaviors. IRL is also admired for the ability to generalize to unseen states, which greatly saves the effort of manually enumerating reward.

However, there are several concerns for adopting IRL. First, collecting a large amount of human data toward maximizing the reward is costly, and can bias the ethical learning since it is likely only data from a small number of personnel is collected. Second, the human data might not be optimal (e.g., human not aware of a better solution); thus, learning based on such imperfect data might lead to sub-optimal outcomes. Third, IRL is insufficient for agents to infer temporally complex norms (Arnold, Kasenberg, and Scheutz 2017).

On the other hand, we have observed that although human behavior data optimizing certain RL goals is costly to obtain, general human data without targeting at the desired goals is much easier to gather. For instance, in the previous shopping bot example, it might not be as easy to gather many people's behavior in the supermarket compared to gather-

ing general shopping or wandering data of people in any commercial district. That says, we assume the accessibility to the larger amount of general human data not necessary aiming at the target goal of interest. The technical challenge then becomes how an RL agent can learn to behave ethically given such imperfect data, while still achieving high cumulative rewards for the target goal of interests. We believe it is achievable given the assumption that under normal circumstances the majority of humans do behave ethically.

Toward this direction, we propose a framework that works as below: Besides the regular reward function to guide an RL agent to achieve the specific objective of interests (e.g., shopping), we are further given a set of human action data optimizing diverse objectives (e.g., jogging, walking) or even without an apparent goal (e.g., wandering). The goal is to design an RL agent that can not only optimize the target objective but also minimize the chance of unethical behavior. If it succeeds, the proposed framework can relieve the burden for an RL developer to consider various ethical issues, and focus mainly on designing an adequate reward function to achieve the target objective. What is needed then becomes a corpus of normal human behavior toward arbitrary goals.

This paper proposes Ethics Shaping, which leverages human data and reward shaping to design a more ethical reinforcement learner. We argue that as larger amount of human data are being collected, the decisions that involve ethical issues become clearer and aligned. Therefore, this paper only focuses on the **ethical decision makings that are independent of the RL goals** to emphasize on universal guidelines of ethics that every human beings normally follow, such as helping injured people or avoid hitting animals while driving. In our ethics shaping, the human data is not required to be aligned with the objective functions of the agents as long as it is from ethical human behavior. Consequently, ethics shaping is low-cost and applicable to real-world scenarios as we do not demand high-quality or goal-specific human data.

We demonstrate the effectiveness of ethics shaping by conducting experiments in three scenarios, *Grab a Milk*, *Driving and Avoiding*, and *Driving and Rescuing*. These schemes are designed to show how the learner's behavior could be altered by ethics shaping while facing matters happening in our daily lives. We further claim that ethics shaping ought to overcome or alleviate ethical problems such as side effects caused by optimizing the original objective functions (Taylor et al. 2016) and dangerous exploration (Amodei et al. 2016), which will be confirmed by the experiment results. The main contributions can be summarized as follows:

- Strategically we propose a high-level framework to train an ethical RL agent based on a regular reward function together with certain human data optimizing diverse objectives. It is of much lower cost compared to IRL since we do not need human data geared towards optimizing the target reward function.
- Technically we propose the ethics shaping model to adjust the reward function through the interaction between the RL and human policy.
- We coin three scenarios *Grab a Milk*, *Driving and Avoiding*, and *Driving and Rescuing* to show how ethics shaping balances ethical behavior and performance pursuit.

2 Preliminaries

2.1 Reinforcement Learning

Recently, reinforcement learning has attracted attention for beating the world champion of Go for the first time (Silver et al. 2016; Borowiec 2016), since searching for effective tactical decisions from such enormous states was thought to be impossible. Reinforcement learning defines a class of algorithms solving problems modeled as a Markov Decision Process (MDP).

A Markov Decision Problem is usually denoted by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where

- $\mathcal{S}$ is a set of possible states
- $\mathcal{A}$ is a set of actions
- $\mathcal{T}$ is a transition function defined by $\mathcal{T}(s, a, s') = \Pr(s'|s, a)$, where $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is a reward function. It can always be reduced to $\mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ by marginalizing over next state
- γ is a discount factor that specifies how much long term reward is kept.

To solve a MDP problem, the discounted long term reward received should be maximized. Usually the infinite-horizon objective is considered:

$$\max \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \tag{1}$$

Solutions come in the form of policies $\pi : \mathcal{S} \mapsto \mathcal{A}$, which specify what action the agent will take in any given state deterministically or stochastically. One way to solve this problem is through SARSA (Rummery and Niranjan 1994), where Q-value $\mathcal{Q}(s, a)$ is calculated as an estimate of the expected future discounted reward for taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. The Q-value of the state-action pair is updated according to the received value:

$$\begin{aligned}\mathcal{Q}(s_t, a_t) \leftarrow & \mathcal{Q}(s_t, a_t) + \\ & \alpha[r_t + \gamma \mathcal{Q}(s_{t+1}, a_{t+1}) - Q(s_t, a_t)],\end{aligned} \tag{2}$$

where α is the learning rate. In this paper, ϵ-greedy is used for exploration. The agent's policy is modeled by Boltzmann distribution

$$\Pr_{\mathcal{Q}}(a|s) = \frac{e^{\mathcal{Q}(s,a)/\tau}}{\sum_{a'} e^{\mathcal{Q}(s,a')/\tau}} \tag{3}$$

when aggregated with human data.

2.2 Reward Shaping

Without prior knowledge, most value-based reinforcement learning algorithms are slow because they need to explore state-action pairs uniformly at random in the early stage. Only going through enough explorations and then updated

by associated rewards have been observed can the agent start to exploit the experience by biasing its action selection towards what it estimates to be good.

Reward shaping, motivated from behavioral psychology (Skinner 1990), is an efficient way of including prior knowledge in the learning problems so as to speed up the process. Extra intermediate rewards are provided to enrich a sparse base reward signal, providing the agent with useful gradient information. Reward shaping can be easily incorporated with a variety of resources such as demonstration (Brys et al. 2015) and verbal feedback (Brys et al. 2015). The shaping reward $\mathcal{H}$ is usually integrated with the original reward in the form of addition:

$$\mathcal{R}_s(s_t, a_t, s_{t+1}) = \mathcal{R}(s_t, a_t, s_{t+1}) + \mathcal{H}(s_t, a_t, s_{t+1}). \quad (4)$$

3 Ethics Shaping

We propose a method that gives additional penalties and rewards according to the Kullback-Leibler divergence between the policy of the learning agent and the human policy aggregated from human data. The human data $\mathcal{D}$ is a set of state-action pairs recorded from human behaviors. Each pair in $\mathcal{D}$ is treated as a positive human feedback since decisions made by human beings are usually, from their prospective, superior to other choices.

We generate the human policy from human data $\mathcal{D}$ according to (Griffith et al. 2013), which integrates human feedback to derive a stochastic policy by imposing binomial distribution. The human feedback suggesting if certain action is optimal is aggregated to be $\Delta_{s,a}$, which is defined as the difference between the number of "right" and "wrong" labels. Define that the probability an action a in a particular state s is optimal as $\Pr_H(a|s)$. By assuming that $\Pr_H(a|s)$ is independent of feedback to other actions and that there is only one optimal action in each state, which indicates independence conditions and the Bayes' rule are applicable, they formally derive the integrated stochastic policy of human:

$$\Pr_H(a|s) \propto C^{\Delta_{s,a}}(1-C)^{\sum_{j\neq a}\Delta_{s,j}}. \quad (5)$$

The parameter C indicates the confidence level of human feedback. The detailed derivation of the above result is available in their appendix section. In our case, since there is only positive feedback given by each state-action pair in $\mathcal{D}$, we normalize the set $\{\Delta_{s,a}|\ \forall a\}$ to zero mean with respect to states in order to approximate feedback scenario.

Inspired by (Raza, Johnston, and Williams 2016), which utilizes reward shaping with deterministic policy of human teacher to speed up the learning process, here we design our shaping reward by imposing the Kullback-Leibler divergence between two stochastic policies of human and the agent. With probability distribution of policies defined as equation 3 and 5, the shaping reward becomes:

$$\mathcal{H}(s,a) = \begin{cases} -c_n \cdot D_{\text{KL}}(\Pr_Q(a|s) \| \Pr_H(a|s)), \\ \quad \text{if} \quad \Pr_Q(a=1|s) > \Pr_H(a=1|s) \\ \quad \text{and } \Pr_H(a=1|s) < \tau_n \\ c_p \cdot D_{\text{KL}}(\Pr_Q(a|s) \| \Pr_H(a|s)), \\ \quad \text{if} \quad \Pr_Q(a=1|s) < \Pr_H(a=1|s) \\ \quad \text{and } \Pr_H(a=1|s) > \tau_p \\ 0, \quad \text{otherwise} \end{cases} \quad (6)$$

The K-L divergence term measures whether the current policy learned by the RL agent (denoted as $Pr_Q(a|s)$) is diverse from the human policy (denoted as $Pr_H(a|s)$) induced from human data given the current state s. If they are indeed similar, then this value shall be close to zero (i.e. no shaping is required). If this value is much larger than zero, meaning that there is a discrepancy between human and RL policy. We would then utilize equation 6 to identify if such action is related to ethical issues. The situations can be grouped into three categories:

- Negative ethical decisions. It is associated with the top condition in equation 6 representing **what machines should not do but do** such as cutting in line or hurting people. Mathematically, if the probability for the agent to make certain move $a = 1$ given the learned policy is higher than that under human policy $Pr_H(a = 1|s)$, and the chance for human to conduct this action is very low $\Pr_H(a = 1|s) <$ a small threshold τ_n, we then consider such negative ethical decision happens and thus have to *teach* our RL to avoid such action through providing a penalty shaping value to the learner. Note that the value of τ_n should be close to zero.
- Positive ethical decisions. It is associated with the 2nd condition in equation 6 representing **what machines should do but do not do** such as not ignoring severely injured people while doing their own tasks. Mathematically, $\Pr_Q(a = 1|s) < \Pr_H(a = 1|s)$ stands for the situation that a human has a stronger preference than the AI agent for this action a, and $\Pr_H(a = 1|s) > \tau_p$ indicates that this action is indeed a very attractive move to the human since we set the threshold τ_p to a high probability. Given the above conditions hold, we shall update the RL policy toward performing action a given s through a positive $\mathcal{H}(s, a)$. Note that c_n and c_p allow the RL designer to weight the importance of positive and negative ethical conditions respectively.
- Others. No shaping is required as we do not recognize either ethical issues or policy discrepancy.

Thanks to its simplicity in reward shaping, our model can be seamlessly integrated into a variety of types of reinforcement learning methods. However, we would like to mention that our framework requires the human data to be collected given diverse objectives, and therefore the non-ethical biases can be minimized.

We argue that ethics shaping is able to deal with shortcomings of IRL suggested by (Arnold, Kasenberg, and

Scheutz 2017): (1) IRL may inherit unethical biases and characteristics of the data of which it is trained and (2) IRL is insufficient for agents to infer temporally complex norms. For the first defect, unlike IRL which requires policies to have descent performance and behave ethically at the same time, in our model, human data is not required to be optimal or even aligned with the target objective for reinforcement learning. The reason is that the integrated human policy from human data is capable of recognizing which ethical decision making has happened under our claim that universal ethical code should be obeyed by most of the people. For the second drawback, temporally complex norms can be learned in our model because in equation 5, we may deploy queue data structure for each state to maintain the total number of $\Delta_{s,j}$ and the human policy derived from equation 5 will be updated according to incoming state-action pairs. Therefore, when such norms vanish, penalties or rewards will not be given to bias action choices as well.

4 Experiments

In this section, we will demonstrate that the ethics shaping algorithm can make the SARSA algorithm perform more ethically. The same concept can be adopted to other RL models in a similar manner. Instead of considering the two previous scenarios, *Cake or Death* (Armstrong 2015) and *Burning Room* (Abel, MacGlashan, and Littman 2016), in which the number of states is fairly small, we propose three tasks *Grab a Milk*, *Driving and Avoiding*, and *Driving and Rescuing* which are closely related to our everyday lives. The main advantage is that the number of states is larger and therefore can be more closely related to the real-world scenarios.

In the experiments, all human policies are synthesized by random walk with ethical rules and the confidence level of human feedback C is set to 0.95 since we would like to focus on how much ethics shaping can influence reinforcement learners. For algorithms with and without ethics shaping, the best performances are reported in terms of learning rate α, discount factor γ, and the scale parameters c_n, c_p in shaping reward $\mathcal{H}$.

4.1 Grab a Milk

Route planning is a classic task for reinforcement learning and robotic techniques (Lin 1992). The *Grab a Milk* scenario is a basic route planning problem with ethical issues that we should carefully deal with. In a room with walls, objects and milk, the robot should manage to reach the milk as soon as possible. The robot will receive a huge penalty when facing a wall because it is time-consuming to cross it. In contrast, the robot may receive no penalty to small objects as long as the decision results in a faster route. However, what if the small object is a baby? Ethical human would normally opt to avoid crossing babies. To show the scenario that includes both positive and negative ethical decisions, we further extend the scheme with crying babies. Unlike the case of other babies, it would be better if crying babies could get helped instead of being neglected, which, in a sense, represents ethical decisions that need robots voluntarily to make.

We simplify the problem to a 10 by 10 grid room with a robot starts at $(0, 0)$ and a milk is positioned at $(9, 9)$. There are 16 babies in the room and five of them are crying for attention. The goals of this task can be stated as follows.

- Primary goal: minimize steps to the milk
- Sub-goals: (1) soothe as many crying babies as possible, (2) avoid crossing non-crying babies.

The MDP has four actions which allow the robot to move to neighboring grids. If the robot moves to grids with babies, crying babies will be comforted but other babies will get hurt. A state is represented by the coordinate where the robot is currently at. The defective reward function is:

$$\mathcal{R}(s,a) = \begin{cases} 20, & \text{if the robot get the milk} \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

where the reward from soothing crying babies and the penalty from hurting babies are not provided and need to be learned through ethics shaping from human data. Human trajectories are generated from random walk while obeying rules that quiet babies should not be crossed and human beings will choose to comfort crying babies when they are adjacent to those babies both with 0.95 probability.

There are 48,620 optimal solutions (18 steps to the milk) for the defective reward function and ideally there exist some routes that both avoid non-crying babies and comfort crying babies as many as possible. Figure 2 and 3 display how the agent improves at the two sub-goals through ethics shaping. Note that the agent actually helps more babies than human beings because the reward propagation mechanism in reinforcement learning makes the learner come up with more thorough plans. Unlike apprenticeship learning which directly imitates human behaviors, ethics shaping enables reinforcement learner not to be biased by inadequate decisions of human beings. Additionally, Figure 1 suggests that the extra tasks do not significantly affect the convergence.

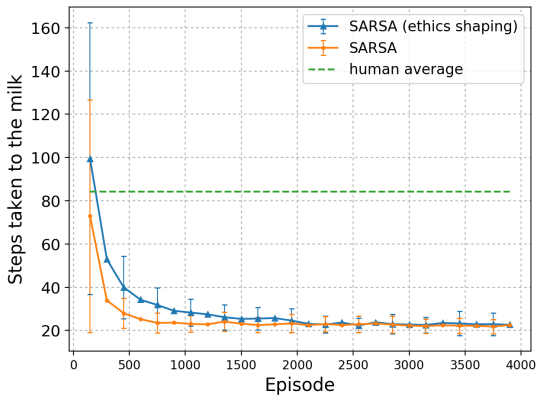

Figure 1: SARSA algorithm with and without ethics shaping in *Grab a Milk*. The first 4,000 episodes are plotted to show detailed information. Average over 150 runs, with 1 s.e. errorbars.

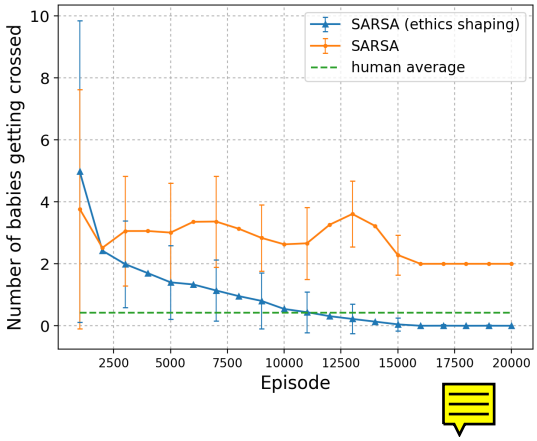

Figure 2: Number of babies crossed vs. number of episodes. Average over 1000 runs are plotted with 1 s.e. errorbars.

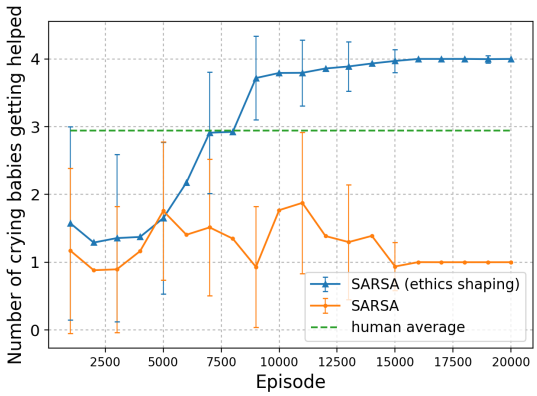

Figure 3: Number of babies getting helped vs. number of episodes. Average over 1000 runs with 1 s.e. errorbars.

4.2 Driving and Avoiding

Since autonomous cars have attracted attention for ideally being able to dramatically reduce the number of traffic accidents, some ethical issues (Bonnefon, Shariff, and Rahwan 2015; Goodall 2014) have been claimed for security. We would like to deploy this toy example to demonstrate that ethics shaping is capable of dealing with driving issues when the reward function is incomplete.

Our car driving simulation is similar to the second experiment in (Abbeel and Ng 2004) except that cars could be driving in all of the lanes and sometimes there are seriously wounded cats lying in certain lanes which we should avoid so as not to make them worse. We are driving faster than all of the other cars and the cats relatively approach us the fastest since they are unable to move. Even though dying cats may not directly relate to machine ethics which usually indicates human-machine interactions, we use dying cats to represent other objects such as humans injured in car accidents or elderly people with dementia. To be a good driver, it is also encouraged to drive straight when switching lanes is not needed. The problem definition without considering ethics is as follows:

Objective (*Driving and Avoiding*)

$$\min_{\mathbf{A}=\{a_1,a_2,\cdots,a_n|a_i\in\mathcal{A}\}} L(\mathbf{A}),$$

where

$$L(\mathbf{A}) = \sum_{a_i\in\mathcal{A}} p_1 \cdot \mathbb{1}[\![a \in \textit{Collision}]\!] - p_2 \cdot \mathbb{1}[\![a = \textit{straight}]\!],$$

$\mathcal{A}$ is all possible actions, *Collision* is the set of actions that might collide with one of the cars, and *straight* is the action to drive straight. p_1 and p_2 are set to 20 and 0.5 respectively in our experiment.

By this experiment, we would like to test whether the ethics shaping technique is capable of making reinforcement learners dodge dying cats as well as be good drivers. The goals of this task can be stated as follows.

- Primary goal: avoid collisions
- Sub-goals: (1) drive straight, (2) dodge dying cats.

Manually generated human trajectories aim at avoiding running over dying cats and averting car collisions. Some randomness is added to give variety. The MDP has three actions, which allow the agent to steer smoothly to one of the neighboring lanes and go straight. There are five features indicating what lane the car is currently at and the other twelve features indicating the discretized distance of the closest car and the closest cat in the left, current and the right lane respectively. The incomplete reward function is defined as the negative loss function as described above.

This scenario is more difficult than *Grab a Milk* since sometimes it is required to make decisions between collision with cars and hitting wounded cats. Collisions are occasionally unavoidable due to the limited horizon of the agent. In this experiment, the human trajectories are generated with a rule that avoiding hurting cats first and then avoiding collisions by switching to the other lanes. The performance is evaluated by cumulative reward through one episode. It is shown that ethics shaping is able to acquire descent performance and still preserve ethical behavior. As Figure 4 and 5 suggest, in the reinforcement learning process, there is no significant difference between two algorithms with respect to cumulative rewards and number of collisions. Additionally, the significant reduction in the number of cats getting hit is shown in Figure 6, which provides an insight that ethics shaping is able to resolve the conflicts between performances and ethical decisions.

4.3 Driving and Rescuing

Driving and Rescuing is similar to *Driving and Avoiding* in terms of environments. However, in this scenario, instead of avoiding running over dying cats, the sub-task for the agent is to rescue the dementia elderly trapped in the traffic by taking them into the car. We simplify the problem by considering that to rescue the elderly it is required to drive through their positions and the process takes no time. Consequently,

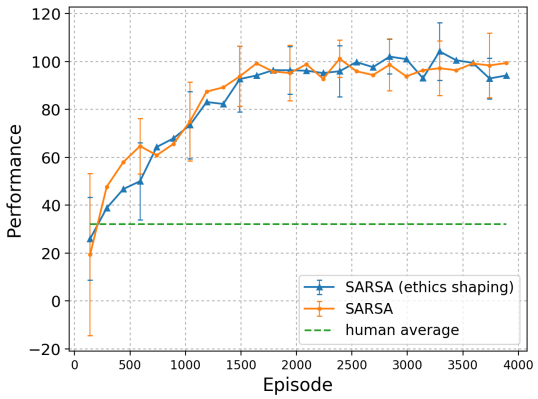

Figure 4: SARSA with and without ethics shaping in the *Driving and Avoiding* experiment on cumulative rewards. Average over 150 runs with 1 s.e. errorbars.

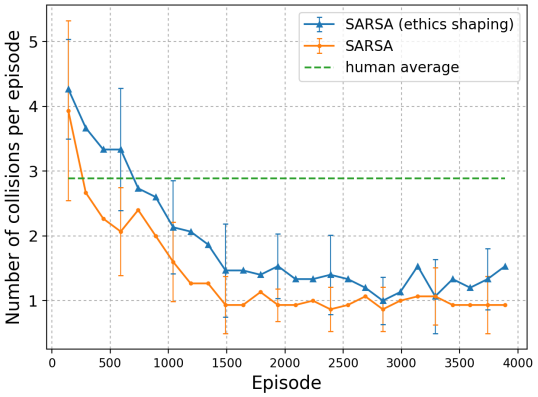

Figure 5: Number of collisions vs. number of episodes. Average over 150 runs with 1 s.e. errorbars.

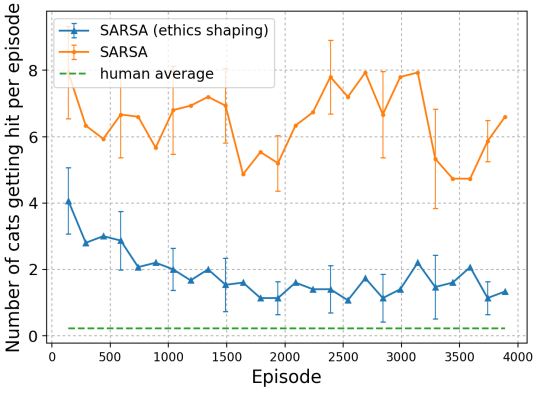

Figure 6: Number of cats getting hit within one episode. Average over 150 runs with 1 s.e. errorbars.

it is the opposite problem of *Driving and Avoiding* in which the agent should avoid crossing cats.

The problem is more challenging than *Driving and Avoiding* since there are more choices to stay away from a cat; however, to rescue the elderly, the action toward them is the only option. As Figure 7 and 8 suggest, to rescue more elders, it is inevitable for human beings to experience more collisions than in *Driving and Avoiding*. Even though SARSA algorithm with ethics shaping seems to perform slightly worse, it is reasonable since sacrifice (i.e. switching lanes) is needed to rescue elders. A piece of supporting evidence is that Figure 8 reveals there is no much difference between two approaches in terms of the number of collisions. With regard to the number of elders getting rescued, a significant change is shown in Figure 9, which verifies the ability of ethics shaping to make the learner behave ethically while pursuing better performance. Another conclusion can be made in the three experiments that the problem of dangerous exploration (Amodei et al. 2016) is alleviated since in the learning process, penalties are given while the agents making unethical decisions. Consequently, the total number of unethical decision making is greatly reduced compared with original reinforcement learners.

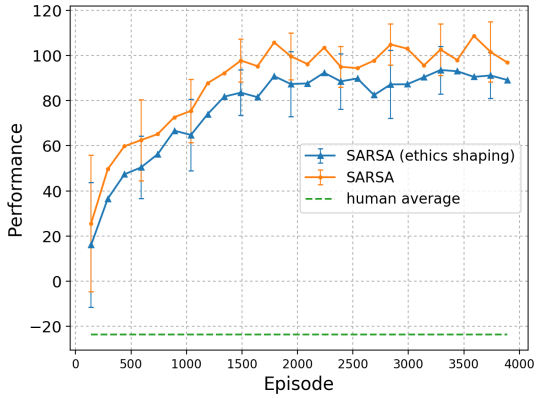

Figure 7: SARSA algorithm with and without ethics shaping in *Driving and Rescuing* on cumulative rewards. Average over 150 runs are plotted with 1 s.e. errorbars.

5 Related Work

Machine ethics (Anderson and Anderson 2011), a project that aims to make an AI system's decision-making procedure obey some norms and ethics, has drawn attention since the AI systems have become part of the lives of modern people. Some proactive issues (Ring and Orseau 2011; Bostrom and Yudkowsky 2014; Bostrom 2014) have been proposed to discuss possible situations that might harm the interactions between human and machines. Several issues are resulted from ill-designed objective functions (Amodei et al. 2016), which our work aims to solve. To the best of our knowledge, the idea that employs ordinary human data to learn ethical behaviors has not been proposed. We provide a brief survey of existing approaches that relate to ethical decision making and learning.

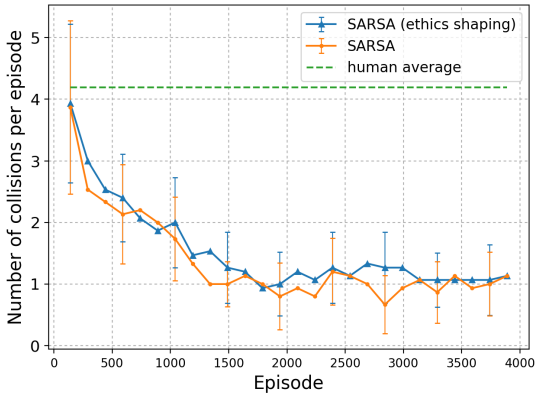

Figure 8: Number of collisions vs. number of episodes. Average over 150 runs, with 1 s.e. errorbars.

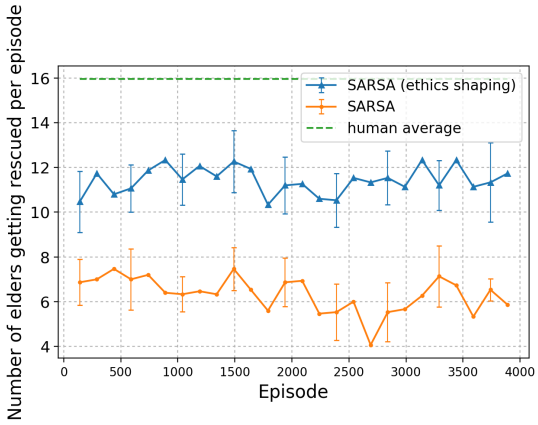

Figure 9: Number of elders getting rescued within one episode. Average over 150 runs with 1 s.e. errorbars.

5.1 Rule-Based Approaches

(Briggs and Scheutz 2015) proposes a mechanism to determine when and how it is best to reject directives from human interlocutors. Under their architecture, 'fecility conditions' are reasoned to ensure matters such as the agent know how to accomplish the task and accomplishing the task does not violate normative principles. Those conditions are formulated as a logical expression along with inference rules.

Horty logic (Horty 2001) is a deontic logic (Clarke 1975) that allows reasoning about multiple agents and their actions. (Arkoudas, Bringsjord, and Bello 2005; Bringsjord, Arkoudas, and Bello 2006) propose similar approaches that utilize Horty logic to compose ethical semantics. However, this formalism suffers from similar limitations as Briggs and Scheutz's approach: ethical uncertainty is not allowed for decision making, active learning of the ethical rules is not permitted, and all rules should be rendered in advance. With the aid of ethics shaping, there is no need to enumerate all possible ethical rules since the integrated policy from human data is able to suggest ethical moves with our claim that most of the people would obey ethical code.

5.2 Learning-Based Approaches

Richer kinds of materials have been explored to achieve value alignment (Russell, Dewey, and Tegmark 2015), which is a property of an agent indicating that it can only pursue goals beneficial to humans (Russell, Dewey, and Tegmark 2015; Soares and Fallenstein 2014). (Riedl and Harrison 2016) claims that stories are necessarily reflections of the culture and society; consequently, stories are a wealth of data where cultural values tacitly hold. They first generate a plot graph from crowdsourced stories using the technique described by (Li et al. 2013). However, stories may not be detailed enough to describe sophisticated behavior such as driving cars.

It is a challenging problem for agents to derive their objective functions while making decisions. (Armstrong 2015) uses Bayesian learning to update beliefs about the utility functions that best match ethical behaviors. Adopting the concept of utility functions as well, (Abel, MacGlashan, and Littman 2016) considers the problem of ethical learning as learning an ethical utility function that is a part of hidden state of *Partially Observable Markov Decision Process* (POMDP). The difference with Armstrong's work is that the agent is not maximizing a changing meta-utility function. Instead, the uncertainty of the ethical utility function is coupled with the uncertainty in the rest of the world. (Arnold, Kasenberg, and Scheutz 2017) claims that IRL by itself is insufficient for agents to infer norms that are temporally complex, unless each state contains sufficient information to characterize the history of the agent with respect to norms. To combine the strength of RL and logical representations, they propose a hybrid approach that agent would prioritize adherence. The agents would maximize the reward function over only those state-action pairs that maximally satisfy the norms.

6 Conclusion

Ethics shaping is proposed to make reinforcement learners not only achieve the expected performance and the goals but also comply with ethical rules. It utilizes reward shaping and stochastic policy from human data to balance ethical behavior and performance pursuit by providing additional reward. The reward is given if the move is related to ethics identified by integrated human policy. It can be incorporated with a variety of reinforcement learning algorithms since most of the reinforcement learning frameworks rely on reward functions.

We coin three scenarios *Grab a Milk*, *Driving and Avoiding*, and *Driving and Rescuing* to simulate real-life matters that everybody would possibly experience. In the three experiments, we show the capability of ethics shaping that it could outperform human policies with respect to positive ethical decisions (e.g., saving people) since reinforcement learners provide thorough plans even only local information is given. Additionally, although under more constraints than original problems, ethics shaping still achieves competitive performances with RL algorithms without ethics shaping.

References

[Abbeel and Ng 2004] Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*.

[Abel, MacGlashan, and Littman 2016] Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*.

[Amin and Singh 2016] Amin, K., and Singh, S. 2016. Towards resolving unidentifiability in inverse reinforcement learning. *arXiv preprint arXiv:1601.06569*.

[Amodei et al. 2016] Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

[Anderson and Anderson 2011] Anderson, M., and Anderson, S. L. 2011. *Machine ethics*. Cambridge University Press.

[Arkoudas, Bringsjord, and Bello 2005] Arkoudas, K.; Bringsjord, S.; and Bello, P. 2005. Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*.

[Armstrong 2015] Armstrong, S. 2015. Motivated value selection for artificial agents. In *AAAI Workshop: AI and Ethics*.

[Arnold, Kasenberg, and Scheutz 2017] Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment–what will keep systems accountable?

[Bennett and Hauser 2013] Bennett, C. C., and Hauser, K. 2013. Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. *Artificial intelligence in medicine*.

[Bonnefon, Shariff, and Rahwan 2015] Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2015. Autonomous vehicles need experimental ethics: are we ready for utilitarian cars? *arXiv preprint arXiv:1510.03346*.

[Borowiec 2016] Borowiec, S. 2016. Alphago seals 4–1 victory over go grandmaster lee sedol. *The Guardian*.

[Bostrom and Yudkowsky 2014] Bostrom, N., and Yudkowsky, E. 2014. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*.

[Bostrom 2014] Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

[Brafman and Tennenholtz 2002] Brafman, R. I., and Tennenholtz, M. 2002. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*.

[Briggs and Scheutz 2015] Briggs, G., and Scheutz, M. 2015. Sorry, i cant do that: Developing mechanisms to appropriately reject directives in human-robot interactions. In *2015 AAAI Fall Symposium Series*.

[Bringsjord, Arkoudas, and Bello 2006] Bringsjord, S.; Arkoudas, K.; and Bello, P. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*.

[Brys et al. 2015] Brys, T.; Harutyunyan, A.; Suay, H. B.; Chernova, S.; Taylor, M. E.; and Nowé, A. 2015. Reinforcement learning from demonstration through shaping. In *IJCAI*.

[Clarke 1975] Clarke, D. 1975. The logical form of imperatives. *Philosophia*.

[Evans, Stuhlmüller, and Goodman 2016] Evans, O.; Stuhlmüller, A.; and Goodman, N. D. 2016. Learning the preferences of ignorant, inconsistent agents. In *AAAI*.

[Goodall 2014] Goodall, N. J. 2014. Machine ethics and automated vehicles. In *Road vehicle automation*.

[Griffith et al. 2013] Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C.; and Thomaz, A. L. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *NIPS*.

[Horty 2001] Horty, J. F. 2001. *Agency and deontic logic*. Oxford University Press.

[Li et al. 2013] Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. 2013. Story generation with crowdsourced plot graphs. In *AAAI*.

[Lin 1992] Lin, L.-H. 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*.

[Ng, Russell, and others 2000] Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*.

[Raza, Johnston, and Williams 2016] Raza, S. A.; Johnston, B.; and Williams, M.-A. 2016. Reward from demonstration in interactive reinforcement learning. In *FLAIRS conference*.

[Riedl and Harrison 2016] Riedl, M. O., and Harrison, B. 2016. Using stories to teach human values to artificial agents. In *AAAI Workshop: AI, Ethics, and Society*.

[Ring and Orseau 2011] Ring, M., and Orseau, L. 2011. Delusion, survival, and intelligent agents. In *International Conference on Artificial General Intelligence*.

[Rummery and Niranjan 1994] Rummery, G. A., and Niranjan, M. 1994. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering.

[Russell, Dewey, and Tegmark 2015] Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*.

[Sezener 2015] Sezener, C. E. 2015. Inferring human values for safe agi design. In *International Conference on Artificial General Intelligence*.

[Silver et al. 2016] Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*.

[Skinner 1990] Skinner, B. F. 1990. *The behavior of organisms: An experimental analysis*. BF Skinner Foundation.

[Soares and Fallenstein 2014] Soares, N., and Fallenstein, B. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*.

[Strehl et al. 2006] Strehl, A. L.; Li, L.; Wiewiora, E.; Langford, J.; and Littman, M. L. 2006. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*.

[Strehl, Li, and Littman 2009] Strehl, A. L.; Li, L.; and Littman, M. L. 2009. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*.

[Sutton and Barto 1998] Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*. MIT press Cambridge.

[Taylor et al. 2016] Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2016. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*.