When Is It Morally Acceptable to Break the Rules? A Preference-Based Approach

Edmond Awad[1], **Sydney Levine**[2], **Andrea Loreggia**[3], **Nicholas Mattei**[4], **Iyad Rahwan**[5], **Francesca Rossi**[6], **Kartik Talamadupula**[7], **Joshua Tenenbaum**[8], and **Max Kleiman-Weiner**[9]

Abstract. Humans make moral judgements about their own actions and the actions of others. Sometimes they make these judgements by following a utilitarian approach, other times they follow simple deontological rules, and yet at other times they find (or simulate) an agreement among the relevant parties. To build machines that behave similarly to humans, or that can work effectively with humans, we must understand how humans make moral judgements. This includes when to use a specific moral approach and how to appropriately switch among the various approaches. We investigate how, why, and when humans decide to break some rules. We study a suite of hypothetical scenarios that describes a person who might break a well established norm and/or rule, and asked human participants to provide a moral judgement of this action. In order to effectively embed moral reasoning capabilities into a machine we model the human moral judgments made in these experiments via a generalization of CP-nets, a common preference formalism in computer science. We describe what is needed to both model the scenarios and the moral decisions, which requires an extension of existing computational models. We discuss how this leads to future research directions in the areas of preference reasoning, planning, and value alignment.

1 Introduction

When we make a moral judgement, depending on the context, we sometimes follow some simple (deontological) rules that have been agreed upon by us or society, or we evaluate the consequences of the possible actions and then we decide (utilitarian), or also we try to find an agreement between the parties involved (contractualism). If we want to build machines that work effectively with us, we need to provide them with some glimpse of our moral judgement methodology. We investigate when humans switch between different frameworks for moral decisions and judgments, and we discuss how to model and possibly embed this switching into a machine.

In particular, we study when humans find it morally acceptable to break simple and generally agreed upon rules. We depart from the typical work in this area which has often focused on high-stakes judgments in extremely uncommon scenarios (e.g., a runaway trolley headed towards innocents) by probing people's moral intuitions in everyday scenarios, such as standing in line to receive a service or to buy some object.

Intuitively, it seems like an easy feat to figure out how to wait in line. In fact, it seems like one simple rule governs the process of waiting in line: each person in line is helped in the order that they arrive (first-in-first-out). If true, navigating a situation that requires waiting in line would simply involve getting in the back of the line and waiting your turn.

However, a few moments' reflection reveals that we can intuitively evaluate all kinds of exceptions to this seemingly-simple rule about waiting in line. For example, say that you are in a deli and have waited in line and ordered a bowl of soup, but just as you are about to begin eating the soup, your spoon falls on the floor. It is probably acceptable for you to ask for a new spoon without waiting in line. Or say that you just want a glass of tap-water. Here, too, usually you are allowed to cut to the front of the line without waiting. Or if you are assisting someone who has just fallen off a bike outside and needs water, you can probably purchase a bottle of water without waiting in line. On the other hand, it is probably not acceptable to cut to the front of the line to order a soda, even though that might delay the line just as much as the person requesting tap water. How do humans figure out when it is acceptable or OK to cut in line and when it's not? And if one decides that it is OK to cut the line, what kind of reasoning and actors does she consider when making this decision? Despite first appearances, figuring out how to wait in line cannot be governed by simple rules.

Contribution. We investigate when humans find it acceptable to break the rules. The goal of these experiments is to understand the ways in which humans evaluate the subtle differences between everyday morally relevant scenarios to make contextualized moral judgements and translate these evaluations into computational models for machine reasoning. To accomplish our task of understanding human moral judgements, we have run experiments that shows a set of scenarios to people. Each scenario is modelled through a set of variables we call *scenario variables* (such as the location, or the reason to ask to cut the line, etc.) and gives rise to a number of descriptive evaluation variables dependent on the scenarios (such as an evaluation of the delay of allowing to cut the line, etc.). We then ask the people to judge whether or not it is OK to cut the line in each scenario. We conjecture that people consider the scenarios and make their moral decision after having estimated the values of the *evaluation variables*. We therefore build a preferences structure, similar to a CP-net, to model the experiments' data by relying on this conjecture. The role of this structure is to allow for a sophisticated reasoning over the scenar-

[1] University of Exeter, UK, email: e.awad@exeter.ac.uk
[2] Massachusetts Institute of Technology and Harvard University, USA, email: smlevine@mit.edu
[3] European University Institute, Italy, email: andrea.loreggia@gmail.com
[4] Tulane University, USA, email: nsmattei@tulane.edu
[5] Massachusetts Institute of Technology, USA, email: irahwan@mit.edu
[6] IBM Research, USA, email: francesca.rossi2@ibm.com
[7] IBM Research, USA, email: krtalamad@us.ibm.com
[8] Massachusetts Institute of Technology, USA, email: josh.tenenbaum@gmail.com
[9] Massachusetts Institute of Technology, USA, email: maxhkw@gmail.com

ios, their evaluations, and the moral decisions. We discuss possible extensions of this model using probabilistic CP-nets as well as probabilistic planning.

2 Philosophical and Psychological Theories of Moral Judgment

The study of normative ethics has typically been divided into three broad camps, based on how moral philosophers think that we should decide which actions are right or wrong. Crudely, consequentialists focus on evaluating *outcomes*, deontologists focus on the use of inviolable *rules*, and contractualists focus on determining an *agreement* to which everyone involved could assent [22].

Theories of moral psychology typically draw on one or two of these ideas to explain the human capacity for moral judgment. Some psychological theories put notions of rules at their center [32, 31] while others (such as dual process theories) incorporate both rules and outcomes [15, 19, 24, 21]. Other theories have pointed towards the usefulness of thinking about agreement-based processes [8, 25]. Yet, to date, no view has attempted to explain how rules, outcomes, and agreement are all integrated in the moral mind.

There have been various attempts by philosophers to unite the three views into a single unified view [35, 20]. Parfit called this unified view a “Triple Theory” since it unified the three strong threads (outcomes, rules, and agreement). We hypothesize that there is an analogous psychological “Triple Theory” that captures the fundamental elements of moral cognition in humans and can also be formalized computationally. Our aim is to describe a unified theory of moral cognition that combines elements of each of the theories of moral philosophy and to build a computational model of this view that would be able to direct the actions of an AI system. This paper begins to explore this possibility and sketches out how this research program may proceed using two distinct computational modeling tools: (generalized) CP-nets and probabilistic planning.

To this end, we collect data from real-life scenarios involving some form of moral judgement and asking several subjects to make the relevant moral decision. We present the subjects scenarios that involve someone attempting to cut in line. This context provides a framework to gather data that could be modeled by a Psychological Triple Theory because rules, outcomes, and agreements are all at play.

We ask subjects to judge whether it is acceptable to cut in line in the cases described. We code each of these cases for obvious instances of a rule violation, where the rule about cutting in line is construed in simple terms (e.g., you must wait in line at a deli if you intend to buy something). We then ask a separate group of subjects to answer questions that assess each scenario on a range of metrics such as how long the cutter would delay the line, the benefit to the cutter, the detriment to the line, and so forth. In this way, we can generate an expected utility calculation for the action. Finally, we ask subjects what would happen if this type of line-cutting always happened, a proxy for whether everyone would agree to allow this person to cut [25].

We study how these three elements can be combined in a computational model to describe and explain our subjects’ judgments about the presented scenarios. First, however, we survey some of the efforts to date on learning and combining normative models of ethics into AI systems. We then turn to the details our specific experiments and analysis of the results. We conclude with a discussion of implications for model reasoning in AI systems.

3 Ethical Reasoning in AI Systems

The idea of teaching machines right from wrong has become an important research topic in both AI [46] and related fields [44]. Our goal is to formalize and understand how to build AI agents that can act in constrained ways that match the normative judgments of humans [37, 27, 28]. There are a number of research projects in this area across computer science including taking sequences of actions in a reactive environment [34], and teaching agents to respond in certain environments [1]. Many of these projects address what is called the *value-alignment* problem [5], that is, the problem of building machines that behave according to values aligned to the human ones [40, 28, 27, 26].

Concerns about the ways in which autonomous decision making systems behave when deployed in the real world are growing. Stakeholders worry about systems achieving goals in ways that are unacceptable according to values and norms of the impacted community, also called “specification gaming” behaviors [37]. Thus, there is a growing need to understand how to constrain the actions of an AI system by providing boundaries within which the system must operate. To tackle this problem, we may take inspiration from humans, who often constrain the decisions and actions they take according to a number of exogenous priorities, be they moral, ethical, religious, or business values [41, 27, 28], and we may want the systems we build to be restricted in their actions by similar principles [5]. The overriding concern is that the agents we construct may not obey these values while maximizing some objective function [42, 37].

Much of the research at the intersection of artificial intelligence and ethics falls under the heading of *machine ethics*, i.e., adding ethics and/or constraints to a particular system’s decision making process [4]. While giving a machine a code of morals or ethics is important, there is still the question of *how to provide the behavioral constraints to the agent*. A popular technique is called the *bottom-up approach*, i.e., teaching a machine what is right and wrong by example [2, 7].

An important aspect for automated decision making systems is ensuring *transparency* and *interpretability*, i.e., being able to see why the system made the choices it did. [43] observe that the Engineering and Physical Science Research Council (EPSRC) Principles of Robotics dictates the implementation of transparency in robotic systems. The authors define transparency in a robotic or autonomous decision making system as “a mechanism to expose the decision making of the robot” .

One prominent example of agents balancing conflicting interests is the case of autonomous cars. There is extensive research from multidisciplinary groups into the questions of when autonomous cars should make lethal decisions [10, 6], how to aggregate societal preferences to make these decisions [33], and how to measure distances between these notions [27, 28, 29]. In a recommender systems setting, a parent or guardian may want the agent to not recommend certain types of movies to children, even if this recommendation could lead to a high reward [7]. Recently, as a compliment to their concrete problems in AI saftey which includes reward hacking and unintended side effects [3], a DeepMind study has compiled a list of specification gaming examples, where different agents game the given specification by behaving in unexpected (and undesired) ways.

4 Experimental Details

In our study, 320 subjects were recruited from Amazon MTURK. Subjects read short vignettes about people waiting in line in three different contexts: at a deli, for a bathroom at a concert venue, and for a security screening at an airport. The vignettes contained descriptions of someone who wanted to cut in line for a range of different reasons. Subjects were randomly assigned to one of two experimental groups: moral judgment or context evaluation. Subjects in the moral judgment group, read all the scenarios (27 total) and were asked to judge whether it was acceptable for the protagonist to cut in line (yes/no).

As an example: "Imagine that there are five people who are waiting in line at a deli to order sandwiches for lunch. There is only one person (the cashier) working at the deli. A customer who is eating soup at the deli dropped his spoon on the floor and needs another one. Is it OK for that person to ask the cashier for a new spoon without waiting in line?"

Subjects in the context evaluation group were randomly assigned to one of three conditions: deli (14 scenarios), bathroom (7 scenarios), or airport (6 scenarios). Subjects evaluated all the vignettes in one context only. These subjects were asked to make factual assessments about each vignette. Subjects answered the following questions about each vignette:

1. **Everyone:** Think about the well-being of all the people in line combined. How are they affected by the person cutting in line? (-50 = a lot worse off; 0 = not affected; 50 = a lot better off)
2. **First Person:** How much worse off/better off is the first person in line? (-50 = a lot worse off; 0 = not affected; 50 = a lot better off)
3. **Middle Person:** How much worse off/better off is a person standing in the middle of the line? (-50 = a lot worse off; 0 = not affected; 50 = a lot better off)
4. **Last Person:** How much worse off/better off is the last person in line? (-50 = a lot worse off; 0 = not affected; 50 = a lot better off)
5. **Cutter:** How much worse off/better off is the person that cut in line? (-50 = a lot worse off; 0 = not affected; 50 = a lot better off)
6. **Universalization:** Think about the person who cut in line. How much worse off/better off would it be for people who come to the deli if everyone who was in this situation cut in line? (-50 = a lot worse off; 0 = not affected; 50 = a lot better off)
7. **Likelihood:** On any given day, how likely is it that this scenario going to happen? (0 = extremely unlikely; 50 = neither likely nor unlikely; 100 = extremely likely)
8. **Delay Time:** How long would the first person in line be delayed? (answer given in minutes and seconds)

We also coded each vignette for the presence of two features: (1) whether the person attempting to cut in line had already waited in line and (2) whether the person attempting to cut had the goal of accessing the main service the line was providing. The main service for the deli line was the sale of an item, for the airport scenario it was security screening, and for the bathroom it was the use of the toilet.

Note that defining what counts as the "main service" being provided to the line is open to interpretation. We think that variation in this interpretation likely impacts subjects' judgments about the acceptability of cutting in line. For instance, someone who views the main service of the deli line as receiving something from the cashier (including receiving something that does not need to be paid for) will likely have a stricter view of who can cut in line compared with someone who thinks that the main function of the line is allowing the deli visitors to buy something. Getting a refill on tap water, for instance, may become less permissible on this revised view. Our characterizations of the main function of the line are rough approximations meant to describe one view that seems commonly held. We leave the potential variance in this factor to future research.

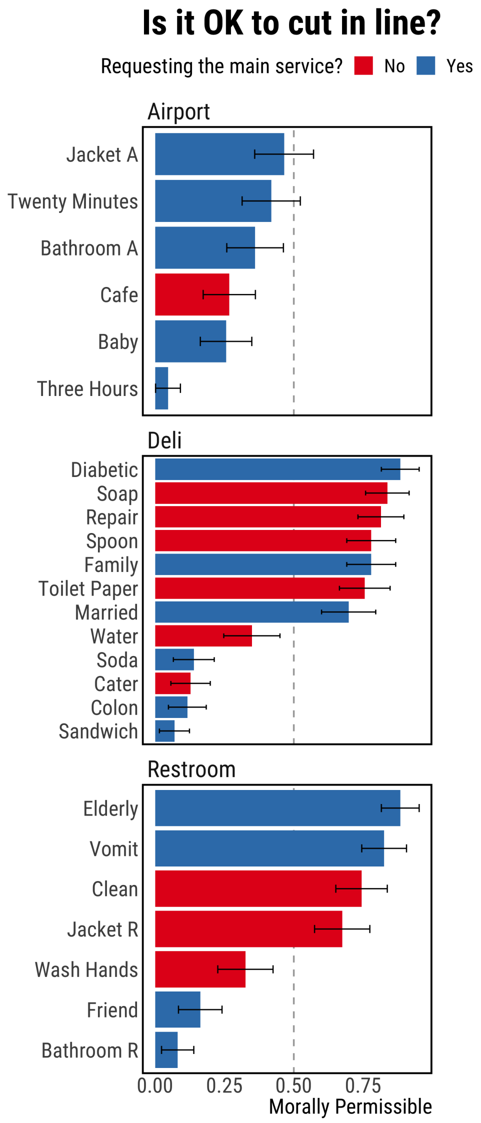

Figure 1: **Moral permissibility of cutting in line.** Error bars are 95% confidence intervals. **Airport.** Five people are waiting in line at an airport. *Jacket A*: someone, who had already waited but had to go get their forgotten jacket, comes back. *Twenty Minutes*: someone whose flight leaves in 20 minutes. *Bathroom A*: someone, who had to leave the line to go to the bathroom, comes back. *Cafe*: someone who works at a cafe inside the airport. *Baby*: Someone with a crying baby. *Three Hours*: someone whose flight leaves in 3 hours. **Deli.** Five people are waiting in line at a deli. *Diabetic*: someone who is diabetic and urgently needs sugar. *Soap*: customer wants to tell the cashier that the bathroom soap needs replacement. *Repair*: oven-repair technician needs to ask the cashier questions so he can fix the oven. *Spoon*: customer who is eating soup dropped his spoon and needs another one. *Family*: a mom and two kids arrive, when the dad is currently placing an order. *Toiletpaper*: customer wants to ask the cashier for toilet paper. *Married*: someone who is married to a customer who is currently placing an order. *Water*: customer who is eating lunch wants a refill on water. *Soda*: customer who is eating lunch wants to buy another soda. *Cater*: someone wants to ask questions about a catering order that he will pick up later. *Colon*: someone who has been fasting for 24 hours in preparation for a colonoscopy and is hungry. *Sandwich*: someone wants to order a sandwich. **Bathroom.** Five people are waiting in line to use a bathroom. *Elderly*: someone is an aid to an elderly person at the front of the line. *Vomit*: someone needs to vomit immediately. *Clean*: someone arrives to clean the bathroom. *Jacket R*: someone forgot their jacket in the bathroom. *Wash Hands*: someone just needs to wash their hands. *Friend*: someone is a friend of someone at the front of the line. *Bathroom R*: someone needs to use the bathroom.

5 Modelling and Reasoning with Preferences

According to Sen [41], moral judgements are a form of preferences, driven by moral reasoning. The issue of modelling and reasoning with preferences in an AI system has been the subject of a very active research area for many years, that produced many frameworks to deal with preferences and embed them into an AI decision making system. Different frameworks differ on properties related, for example, to expressivity, computational complexity, and easiness of preference elicitation.

Conditional Preference networks (CP-nets) are a graphical model for compactly representing conditional and qualitative preferences [11]. CP-nets are comprised of sets of *ceteris paribus* preference statements (cp-statements). For instance, the cp-statement, *"I prefer red wine to white wine if meat is served,"* asserts that, given two meals that differ *only* in the kind of wine served *and* both containing meat, the meal with red wine is preferable to the meal with white wine. CP-nets have been extensively used in the preference reasoning preference learning and social choice literature as a formalism for working with qualitative preferences [16, 38, 13]. CP-nets have even been used to compose web services [45] and other decision aid systems [36].

Formally, a CP-net has a set of features (or variables) $F = \{X_1, \ldots, X_n\}$ with finite domains $D(X_1), \ldots, D(X_n)$. For each feature X_i, we are given a set of *parent* features $Pa(X_i)$ that can affect the preferences over the values of X_i. This defines a *dependency graph* in which each node X_i has $Pa(X_i)$ as its immediate predecessors. An *acyclic* CP-net is one in which the dependency graph is acyclic. Given this structural dependency information among a CP-net's variables, one needs to specify the preference over the values of each variable X_i for *each complete assignment* to the parent variables, $Pa(X_i)$. This preference is assumed to take the form of a total or partial order over $D(X_i)$. A cp-statement for some feature X_i that has parents $Pa(X_i) = \{x_1, \ldots, x_n\}$ and domain $D(X_i) = \{a_1, \ldots, a_m\}$ is a total ordering over $D(X_i)$. The set of cp-statements regarding a certain variable X_i is called the cp-table for X_i.

The semantics of CP-nets depends on the notion of a *worsening flip*: a change in the value of a variable to a less preferred value according to the cp-statement for that variable. For example, in the CP-net above, passing from $abcd$ to $\overline{a}bcd$ is a worsening flip since c is better than $\overline{c}$ given a and b. One outcome α is *preferred to* or *dominates* another outcome β (written $\alpha \succ \beta$) if and only if there is a chain of worsening flips from α to β. This definition induces a preorder over the outcomes, which is a partial order if the CP-net is acyclic [11].The complexity of dominance and consistency testing in CP-nets is an area of active study in preference reasoning [18, 38]. Finding the optimal outcome of a CP-net is NP-hard [11] in general but can be found in polynomial time for acyclic CP-nets by assigning the most preferred value for each cp-table. Indeed, acyclic CP-nets induce a lattice over the outcomes.

6 Handling Morality Driven Preferences

In a standard CP-net, there is only one kind of variable: those needed to express preferences. There is no ability to describe the context in which a preference-based decision making process takes place, nor to model other auxiliary variables that may be needed, or useful, to declare our own preferences. In some sense, a CP-net is a useful tool only when it is clear what the context is, and if no reasoning on the context is needed in order to state the preferences (or such reasoning takes place outside the CP-net formalism).

6.1 Thinking Fast and Slow about Morality

We start by taking inspiration from the ideas of Daniel Kahneman [23] and his description of the two systems, System 1 and System 2, that are relevant for human decision making. The idea is that when making decisions humans employ two different systems: the first is reactive and makes immediate responses that are hard and fast, but sometimes wrong. In other cases a longer term thought process is invoked and humans think about all the factors that go into a decision before making it. There have been proposals to extend this model into reasoning and preference systems in computer science in a principled and exact way [39].

For normative ethical judgements, one could interpret System 1 and System 2 in the context of deontological reasoning and utilitarian reasoning [22]. Informally, System 1 applies some hard and fast rule and does not consider the scenario or try to evaluate the complexities of the current scenario. In contrast, System 2 acts more like a utilitarian reasoner, it attempts to quantify and apply a logic to the scenario. We take inspiration from this idea and attempt to model this reasoning process in the CP-net formalism so that we can embed it within a machine [39], to allow the machine to reason both fast and slow about ethical principles [27]. Our motivation is to extend the semantics of CP-nets in order to model both the snap judgements that do not take into account the particularities of the scenario as well as provide the ability to reason about these details, if necessary.

6.2 Extending CP-nets for Morality Driven Preferences

Analyzing the scenarios of our vignettes and our conjecture of how the subjects reason about the scenarios and then respond to the single preference question (whether somebody should be allowed to cut the line or not), we propose a generalization of the CP-net formalism to handle variables associated with the context. We propose extending the formalism with a set of *scenario variables (SVs)* to define a decision making context over which there is no preference to be stated and a set of *evaluation variables (EVs)* to model the introspection that takes place in the subjects' minds while reasoning over the given context to decide their preference over the standard *preference variables (PVs).*

Figure 2 describes this generalized preference framework visually. We clarify the necessary extra variables as:

1. **Scenario Variables.** A set of variables that describe the context, such as location, whether or not the agent had already waited in line, whether or not the agent was using the main function of the line, and the size of the line. In addition, we need a variable to specify the main reason or motivation for cutting the line. We observe that the agent does not have the ability to set values for these variables, nor does the agent have preferences over their values, as these values are set by the environment or context within which the decision is taking place. These variables do not depend on any other variable (that is, there is no incoming dependency arrow), meaning that this is part of the input to a decision making AI system.
2. **Evaluation Variables.** A set of variables that a person (or an AI system) considers (and estimates the value of) to reason about the given scenario. These are, for example, the well-being of the first in line, the well-being of the cutter, and others as discussed in the

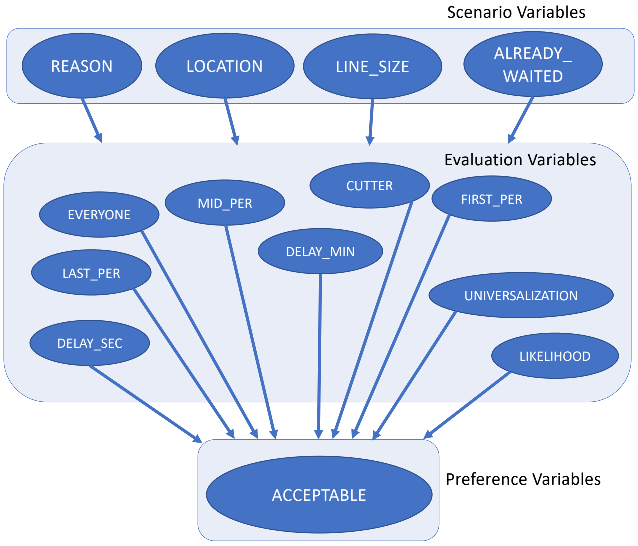

Figure 2: Our conjectured model that blends the intuition of a CP-net with the idea of Scenario and Evaluation Variables. While individuals cannot set or have preferences over the Scenario Variables, they will possess their own subjective evaluations over the Evaluation Variables given a setting to the Scenario Variables. Given both the Scenario Variables and the Evaluation Variables, the agent can then decide on a preference over the single Preference Variable.

experimental details section. In our experiments we have 8 of such variables. These variables have a real valued range as a domain and the user selects one point in the range, which represents her estimate for that variable's value. However, no preferences for the values of these variables is required. All the evaluation variables depend on the scenario variables. This follows our conjecture that people need to examine the specific scenario in order to start an introspection phase in which the identify the evaluation variables and estimate a value for them.

3. **Preference Variables.** In the setting under study the agent really only expresses preference over a single value, that models whether or not, given both the values of the Scenario Variables and the values for the Evaluation Variables, it is acceptable to cut the line. The single preference variable depends on the evaluation variables. This again follow from the conjecture that a person needs to first perform a level of consequentialist introspection in order to decide whether the deontological rule, that states that a line cannot be cut, can be violated.

As noted above, CP-nets and their variants (like probabilistic CP-nets [14]), allow only for preference variables, and there is no option for creating a dependency between the preference variables to scenario and context variables. We envision a three-layer generalization where, as shown in Figure 2, the single preference variable depends on the evaluation variables, which in turn depend on the scenario variables. However, a finer grained analysis may show that there are evaluation variables that do not depend on the scenario variables. For example, the evaluation variable that has to do with the likelihood of the event happening does not have any relationship with whether or not the cutter is concerned with the main function of the line.

6.3 Data Analysis

We asked questions about both scenario variables (SVs), and evaluation variables (EVs) in our surveys. For the two SVs: $LOCATION = \{DELI, BATHROOM, AIRPORT\}$ and $REASON$, whose values are detailed in Section 4, e.g., possible reasons to cut the line to ask for a new spoon without waiting. There are 25 different reasons for cutting, however, only different subsets of them are used for each location as they are context specific. Individuals answered nine evaluation questions for each vignette and we associate each of these with an EV. In what follows, we refer to each of EVs with labels reported in Section 4, where $DELAY\_MIN$ and $DELAY\_SEC$ correspond to answer in minutes and seconds to the eighth question about delay time.

Using the results of our surveys, we want to understand which SVs influence the way individuals respond to EVs. If we can find a statistical dependency, then we say that EVs depend on SVs and validate our model in Figure 2. We also checked whether SVs influence the PV. To test for dependency, we run Wilcoxon signed-rank tests [30] (a non-parametric t-test) for comparing paired data samples from the evaluations of individuals. Using this we can evaluate whether or not we can reject the following three null hypotheses (NH):

1. NH1: location does not affect EVs;
2. NH2: reason does not affect EVs;
3. NH3: location does not affect the PV.

Based on the findings we can construct a partial graph of the dependencies between variables as depicted in Figure 3. We are able, to varying degrees, reject all three of the NHs listed above. Below, failing to reject means that a (set of) pairwise tests have $p\_value \geq 0.01$. Specifically:

NH1 partially rejected: some EVs are influenced by the location and some others are not, as depicted in Figure 3. It is interesting to notice that there are a set of variables that seem to be independent of location. This implies that no matter what location, the value of the EV does not change.

NH2 rejected: all the EVs are influenced by the reason. This is not surprising, as we were expecting that individuals evaluate the scenario differently based on what is happening in the vignette.

NH3 partially rejected: we selected four reasons for each location (since there were different numbers per location), aggregated these, and compared the response to the PV. From this we can re-

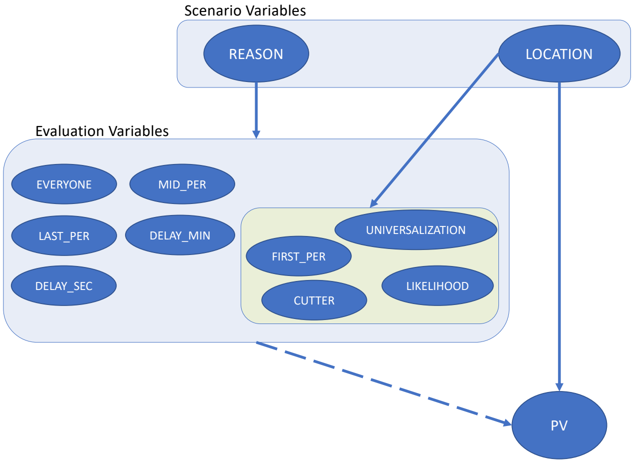

Figure 3: Inferred dependency graph from our data. SVs influence the way individuals evaluate each scenario and make a decision. For the sake of readability we group evaluation variables based whether they depend or not from a scenario variable. This to reduce the number or arrows. The dashed line is a supposed dependency we cannot derive from actual data. Given the SVs and the EVs, the agent can then decide on a preference over the single PV.

ject NH3 for all pairs except Deli and Airport. This indicates that in some cases location may be sufficient to evaluate the vignette and make a decision.

Unfortunately our survey instruments were not designed to show a direct influence of EVs on the PV. We plan to run additional experiments to collect data which help us understand this relationship.

7 On-Going and Future Work

Generalizing CP-nets to Model Moral Preferences. We currently define generalized CP-nets in a way that is consistent with classical CP-nets and probabilistic CP-nets. The aim is to understand how to use a (generalized) preference structure to effectively learn and reason with morality-driven preferences, and to embed them into an AI system.

Prescriptive Plans Based on Moral Preferences. The AI research community has not only been active in understanding how to make single decisions based on preferences, but also on creating plans, consisting of sequences of actions, that would respect or follow certain preferences [12]. This work can be exploited to extend the use of the moral preferences discussed in this paper into more prescriptive AI techniques such as automated planning [9]. Although prior efforts from the planning perspective all investigate the generation of plans that take into account pre-specified utilitarian preferences, the question of where those utilities and preferences manifest from has not been addressed very adequately so far. We are currently actively investigating methods that seek to use the data collected in this work to automatically generate preferences in the notation used by planning formalisms [17]. The generation of such preferences will in turn enable us to generate prescriptive plans for agents or systems that conform to the moral standards of that agent or system. Specifically, we will transform the problem from a classification-based setting into a generative model, and then present plan (action) alternatives that agents can choose from. To situate this in the context of the current question of study: this extension would enable us to move from determining whether it was acceptable to break a rule, to generating ways to do so that are most in accordance with some preference and cost function that takes moral obligations into account.

8 Conclusion

We have taken a first step to model and understand the question of when it is morally acceptable to break rules; and how, why, and when humans decide to do so. We constructed and studied a suite of hypothetical scenarios relating to this question, and collated human moral judgements on these scenarios. We showed that existing structures in the preference reasoning literature are insufficient for this task. We look towards extending this into other established areas of AI research.

REFERENCES

[1] Shani Alkoby, Avilash Rath, and Peter Stone, 'Teaching social behavior through human reinforcement for ad hoc teamwork-the STAR framework', in *Proc. of the 18th AAMAS*, (2019).

[2] 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).

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, 'Concrete problems in AI safety', *arXiv preprint arXiv:1606.06565*, (2016).

[4] Michael Anderson and Susan Leigh Anderson, *Machine Ethics*, Cambridge University Press, 2011.

[5] T. Arnold, Thomas, D. Kasenberg, and M. Scheutzs, 'Value alignment or misalignment - what will keep systems accountable?', in *AI, Ethics, and Society, Papers from the 2017 AAAI Workshop*, (2017).

[6] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, 'The moral machine experiment', *Nature*, **563**(7729), 59, (2018).

[7] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi, 'Incorporating behavioral constraints in online AI systems', in *Proc. of the 33rd AAAI*, (2019).

[8] Nicolas Baumard, Jean-Baptiste André, and Dan Sperber, 'A mutualistic approach to morality: The evolution of fairness by partner choice', *Behavioral and Brain Sciences*, **36**(1), 59–78, (2013).

[9] J Benton, Amanda Coles, and Andrew Coles, 'Temporal planning with preferences and time-dependent continuous costs', in *Proc. 22nd ICAPS*, (2012).

[10] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, 'The social dilemma of autonomous vehicles', *Science*, **352**(6293), 1573–1576, (2016).

[11] C. Boutilier, R. Brafman, C. Domshlak, H.H. Hoos, and D. Poole, 'CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements', *Journal of Artificial Intelligence Research*, **21**, 135–191, (2004).

[12] Ronen I Brafman and Yuri Chernyavsky, 'Planning with goal preferences and constraints.', in *ICAPS*, pp. 182–191, (2005).

[13] *Handbook of Computational Social Choice*, eds., F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, Cambridge University Press, 2016.

[14] C. Cornelio, J. Goldsmith, N. Mattei, F. Rossi, and K.B. Venable, 'Updates and uncertainty in CP-nets', in *Proc. of the 26th AUSAI*, (2013).

[15] Fiery Cushman, 'Action, outcome, and value: A dual-system framework for morality', *Personality and social psychology review*, **17**(3), 273–292, (2013).

[16] C. Domshlak, E. Hüllermeier, S. Kaci, and H. Prade, 'Preferences in AI: An overview', *Artificial Intelligence*, **175**(7), 1037–1052, (2011).

[17] Alfonso Gerevini and Derek Long, 'Plan constraints and preferences in pddl3', *ICAPS*, 7, (2006).

[18] J. Goldsmith, J. Lang, M. Truszczyński, and N. Wilson, 'The computational complexity of dominance and consistency in CP-nets', *Journal of Artificial Intelligence Research*, **33**(1), 403–432, (2008).

[19] Joshua David Greene, *Moral tribes: Emotion, reason, and the gap between us and them*, Penguin, 2014.

[20] Richard Mervyn Hare, Richard Mervyn Hare, Richard Mervyn Hare Hare, and Richard M Hare, *Moral thinking: Its levels, method, and point*, Oxford: Clarendon Press; New York: Oxford University Press, 1981.

[21] Keith J Holyoak and Derek Powell, 'Deontological coherence: A framework for commonsense moral reasoning.', *Psychological Bulletin*, **142**(11), 1179, (2016).

[22] Shelly Kagan, *Normative ethics*, Routledge, 2018.

[23] Daniel Kahneman, *Thinking, fast and slow*, Macmillan, 2011.

[24] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B Tenenbaum, 'Inference of intention and permissibility in moral decision making.', in *CogSci*. Citeseer, (2015).

[25] Sydney Levine, Max Kleiman-Weiner, Fiery Cushman, and Joshua B Tenenbaum, 'What if everyone did that? universalization as a mechanism of moral judgment.', in *CogSci*. Citeseer, (2019).

[26] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable, 'On the distance between cp-nets', in *Proc. of the 17th AAMAS*, pp. 955–963. International Foundation for Autonomous Agents and Multiagent Systems, (2018).

[27] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable, 'Preferences and ethical principles in decision making', in *Proc. 1st AIES*, (2018).

[28] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable, 'Value alignment via tractable preference distance', in *Artificial Intelligence Safety and Security*, ed., R. V. Yampolskiy, chapter 16, CRC Press, (2018).

[29] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable, 'Metric learning for value alignment', in *Proc. of the of AISafety@IJCAI 2019*, volume 2419 of *CEUR Workshop Proceedings*. CEUR-WS.org, (2019).

[30] Henry B Mann and Donald R Whitney, 'On a test of whether one of two random variables is stochastically larger than the other', *The annals of mathematical statistics*, 50–60, (1947).

[31] John Mikhail, *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*, Cambridge University Press, 2011.

[32] Shaun Nichols and Ron Mallon, 'Moral dilemmas and moral rules', *Cognition*, **100**(3), 530–542, (2006).

[33] R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia, 'A voting-based system for ethical decision making', in *Proc. of the 32nd AAAI*, (2017).

[34] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi, 'Teaching ai agents ethical values using reinforcement learning and policy orchestration', in *Proc. of the 28th IJCAI*, (2019).

[35] Derek Parfit, *On what matters: volume one*, volume 1, Oxford University Press, 2011.

[36] P. Pu, B. Faltings, L. Chen, J. Zhang, and P. Viappiani, 'Usability guidelines for product recommenders based on example critiquing research', in *Recommender Systems Handbook*, eds., F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, 511–545, Springer, (2011).

[37] F. Rossi and N. Mattei, 'Building ethically bounded AI', in *Proc. of the 33rd AAAI*, (2019).

[38] F. Rossi, K.B. Venable, and T. Walsh, *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*, Morgan and Claypool, 2011.

[39] Francesca Rossi and Andrea Loreggia, 'Preferences and ethical priorities: Thinking fast and slow in AI', in *Proc. of the 18th AAMAS*, pp. 3–4, (2019).

[40] Stuart Russell, Daniel Dewey, and Max Tegmark, 'Research priorities for robust and beneficial artificial intelligence', *AI Magazine*, **36**(4), 105–114, (2015).

[41] Amartya Sen, 'Choice, ordering and morality', in *Practical Reason*, ed., S. Körner, Blackwell, Oxford, (1974).

[42] T. Simonite, 'When bots teach themselves to cheat', *Wired Magazine*, (August 2018).

[43] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson, 'Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots', in *AISB Workshop on Principles of Robotics*. University of Bath, (2016).

[44] Wendell Wallach and Colin Allen, *Moral machines: Teaching robots right from wrong*, Oxford University Press, 2008.

[45] Hongbing Wang, Shizhi Shao, Xuan Zhou, Cheng Wan, and Athman Bouguettaya, 'Web service selection with incomplete or inconsistent user preferences', in *Proc. 7th International Conference on Service-Oriented Computing*, 83–98, Springer, (2009).

[46] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang, 'Building ethics into artificial intelligence.', in *Proc. 27th IJCAI*, pp. 5527–5533, (2018).