When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data

Citation: Awad, E., Levine, S., Loreggia, A. et al. When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data. Auton Agent Multi-Agent Syst 38, 35 (2024).

As Published: 10.1007/s10458-024-09667-4

Publisher: Springer Science and Business Media LLC

Persistent URL: https://hdl.handle.net/1721.1/155691

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Autonomous Agents and Multi-Agent Systems (2024) 38:35
https://doi.org/10.1007/s10458-024-09667-4

When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data

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

Accepted: 7 July 2024

Abstract
Constraining the actions of AI systems is one promising way to ensure that these systems behave in a way that is morally acceptable to humans. But constraints alone come with drawbacks as in many AI systems, they are not flexible. If these constraints are too rigid, they can preclude actions that are actually acceptable in certain, contextual situations. Humans, on the other hand, can often decide when a simple and seemingly inflexible rule should actually be overridden based on the context. In this paper, we empirically investigate the way humans make these contextual moral judgements, with the goal of building AI systems that understand when to follow and when to override constraints. We propose a novel and general preference-based graphical model that captures a modification of standard *dual process* theories of moral judgment. We then detail the design, implementation, and results of a study of human participants who judge whether it is acceptable to break a well-established rule: *no cutting in line*. We then develop an instance of our model and compare its performance to that of standard machine learning approaches on the task of predicting the behavior of human participants in the study, showing that our preference-based approach more accurately captures the judgments of human decision-makers. It also provides a flexible method to model the relationship between variables for moral decision-making tasks that can be generalized to other settings.

Keywords Moral constraints · Thinking fast and slow · CP-net · Human judgment · Moral decision-making

1 Introduction

Concerns about the ways AI systems behave when deployed in the real world are growing in the research community and beyond [1–3]. A central worry is that AI systems may achieve their objective functions in ways that are morally unacceptable to those impacted by the decisions, for example, revealing "specification gaming" behaviors [4, 5]. Thus, there is a growing need to understand how to *constrain* the actions of AI systems by providing boundaries within which the system must operate.

Extended author information available on the last page of the article

However, imposing formal constraints within the models used by AI systems presents its own problems. Sometimes, constraints preclude actions that are perfectly acceptable or even desirable. On the other hand, sometimes constraints will permit actions that should be prohibited. These properties are true of laws and rules generally speaking, not just for those constraining AI systems. Consider a rule prohibiting vehicles from entering a public park ("no vehicles in the park") [6] in order to preserve the safe and serene environment that the park provides. A plain reading of this rule might presume that ambulances and wheelchairs are barred from entering, though they should clearly be allowed, but might fail to preclude drones, which are not vehicles *per se*, but potentially as dangerous and disruptive. Constraints on AI systems can fall victim to these two kinds of problems— sometimes being too lenient—finding "loop holes" around constraints while optimizing an objective function [5, 7]—while other times being unnecessarily stringent (for a non-exhaustive list see, e.g., [8]).

But few humans would be confused about the rule banning vehicles from the park. No wheelchair user would wonder if they could enter, and ambulances would not meet their arrival on the scene of an accident with protest. In short, humans are experts at figuring out when constraints can and should be broken. How do humans do this and can we enable AI to do the same?

In recent years, there has been an ever-increasing number of proposals on how and why to constrain the actions of AI systems to ensure their alignment with human values, i.e., to make them "ethically aligned" [1], culminating with both the ACM Statement on Algorithmic Transparency [9] and the National Institute of Standards and Technology AI Risk Management Framework [10]. Within the AI research literature, many techniques for constraint engineering are based on bottom-up, top-down, or hybrid approaches [11–13]. A bottom-up approach involves teaching a machine what is right and wrong by providing examples of "correct" decisions—for instance, providing an autonomous vehicle with demonstrations about how to handle certain traffic patterns [13–15]. In top-down approaches, behavioral guidelines are specified by explicit rules or constraints on the decisions space—for instance, by making a rule that an autonomous vehicle must never strike a human [16, 17]. However, both of these kinds of models will struggle to determine when constraints should be overridden. Top-down approaches will err when the rules or constraints on the system are too general to deal with a particular edge case or unusual circumstance. Bottom-up approaches will err when a case that should be an exception to a rule differs dramatically from anything in the training set.

Our goal is to formalize and understand how to build AI agents that can act in ways that are morally acceptable to humans [4, 18] by taking inspiration from how humans decide what is morally acceptable. In this work, we start from the observation that decisions are linked to preferences: we humans choose one among all possible decisions because we prefer the resulting state-of-affairs over the ones generated by the other decisions. Thus, we can model how an individual makes a decision in a given scenario by modeling her preferences over decisions in that scenario [19].

In this paper, we extend the existing preference framework of CP-nets [20] to capture complex preferences over both the *context* of a decision and the resulting *judgments* about actions. Traditional CP-nets allow for the representation of conditional preferences over options. Our novel extension, *Scenario-Evaluation-Preference Nets (SEP-nets)*, allows for a more complete model of how humans decide when it is permissible to break a moral rule. The aim of using this formalism to describe human moral decision-making is to be able to ultimately embed the human-like process of flexibly overriding

constraints into a machine or to build machines that can better collaborate with humans, knowing how humans would behave.

While there are many formalisms to choose from when modeling preferences, we focus on CP-nets because of their many strengths. In particular, they provide a compact representation of possibly complex domains, e.g., in the cloud service selection [21] where they are employed to represent the preferences of users. Moreover, they have been extended in multiple ways to be easily translated into formal languages [22] or to embed uncertainty [23]. The varied number of applications of CP-nets and their popularity in the academic research literature make this framework and its extensions of particular interest as they provide feasible ways for representing conditional preferences and at the same time leverage a plethora of tools for the development of such frameworks.

Our extension of CP-nets, SEP-nets, captures a modification to standard "dual process" theories of moral judgment. Human decision-making processes (in many domains, not just morality) are sometimes characterized by dual process models of cognition, which posit that humans have both a fast and automatic way of making decisions ("System 1") and a slower, more deliberate method ("System 2") [24, 25]. Dual process models of *moral* cognition follow this mold: these theories tend to argue that we make moral judgments using a combination of heuristic-like rules (System 1) and more deliberative utility calculations (System 2) [26, 27]. We argue for a modification of this standard account, proposing that System 1 processes support rule-based moral judgment (as is widely agreed upon in the literature) and that System 2 processes can enable *moral rule flexibility*. Specifically, humans use System 2 thinking to figure out when a rule should be overridden.

To assess our proposal, we run a study in which participants make judgments about breaking a simple and well-known socio-moral rule, specifically, "no cutting in line".[1] Despite first appearances, figuring out how to wait in line cannot be governed by a simple rule; on the contrary, we can intuitively evaluate a huge range of exceptions to the rule about line waiting. We perform an extensive analysis of the data to uncover possible relationships among the variables used to describe the scenario. We then use the collected data to build an SEP-net, and finally, we compare this SEP-net approach with well-known machine learning models including XGBoost [29] and Support Vector Machines [30] on a prediction task and find that our model outperforms the competitors.

1.1 Contribution

We propose a novel extension of CP-nets we call *Scenario-Evaluation-Preference Nets (SEP-nets)*. This framework is able to represent conditional preferences and take into account a complex deliberation process to compute an outcome. We focus on scenarios when a socio-moral rule should be broken, specifically, when it is morally acceptable to cut in line. To support our model, we collect data from human participants, asking them to make judgments about when it is permissible to cut in line. We perform an extensive data analysis and show how our framework can be used to model the moral judgments made by individuals. Our SEP-nets method performs better than most standard machine learning classification algorithms. We suspect that many social and moral rules are supported by

[1] We depart from the typical work in this area which has often focused on understanding judgments in high-stakes, uncommon scenarios, e.g., a runaway trolley headed towards people [28]. Instead, we probe people's moral intuitions about commonplace scenarios to ensure that we are modeling the processes people are using in their day-to-day moral decisions.

generative psychological mechanisms that allow people to make judgments in novel scenarios. We, therefore, believe that the formalism we develop that captures the psychological process that enables people to navigate exceptions to line-waiting rules will be generalizable to many other rules guiding human social life.

1.2 Organization

The rest of the paper is structured as follows. Section 2 details related work in moral psychology as well as in the computational decision-making literature. In Sect. 3 we provide a comprehensive background and formal notation for CP-nets. Section 4 formalizes SEP-nets. Section 5 gives a detailed overview of our experimental setting, data collection and analysis, as well as the empirical evaluation of SEP-nets. Section 6 provides a discussion about the results and Sect. 7 offers conclusions and directions for future work.

2 Related work

The following section provides an overview of the existing research and developments in the field, highlighting key contributions and insights relevant to our study.

2.1 Computational ethics

The vast majority of work characterizing human moral decision-making, both in psychology [31, 32] and experimental philosophy [33, 34] has focused on identifying factors that are relevant to moral judgment (e.g., affect, rules, utility calculations). In this work we focus on an emerging body of work, called "computational ethics" [35], that goes beyond simply identifying these factors, but also seeks to characterize the mechanisms underlying moral judgments in order to embed them in artificial agents [36–42]. This is a critical step in building AI that can produce and interpret human moral judgment [37, 43, 44] in a way that is explainable [45].

One prominent example of agents that try to find a trade-off between maximizing an objective function while respecting ethical constraints is the case of autonomous cars. A range of interdisciplinary research groups have asked how autonomous cars should handle cases where harm to passengers or pedestrians is inevitable [28, 46], how to aggregate societal preferences to make these decisions [47–49], and how to measure distances between these preferences [18, 50]. Similar tensions around ethical constraints arise for recommender systems. A parent or guardian may want the online agent to not recommend certain types of movies to children, even if this recommendation could lead to a high reward [14]. Not being able to specify the appropriate ethical constraints may lead to undesired and unexpected behavior. This is sometimes referred to as "specification gaming" because agents "game" the given specification by behaving in unexpected (and undesired) ways [5].

2.2 Preferences in artificial intelligence

As we noted above, the concepts of decisions and actions are linked to the concept of preferences, and the issue of modeling and reasoning with preferences in an artificial agent has been the subject of research within AI for many years [51]. Several frameworks have been

defined, and their properties studied, for many situations including expressivity [19], computational complexity, and easiness of preference elicitation [52]. The centrality of preferences is true also in the case of moral judgement: we consider a decision more morally acceptable than another one if its impact on others is preferred according to our moral values [53–55]. Therefore, finding a way to model these values, and the corresponding preferences they create, is central to building artificial agents that behave in a way that is aligned to humans values [4, 18].[2]

For instance, Freedman et al. [56] introduce a comprehensive methodology involving human participants to estimate weights for individual profiles in kidney exchanges, ultimately prioritizing patients and donors during organ allocation. This highlights the significant impact of human value judgments on patient prioritization outcomes. WeBuildAI [57] is another illustration, presenting a participatory framework where stakeholders collaboratively build computational models to guide voting-based algorithmic policy decisions. This was demonstrated through a case study involving an on-demand food donation transportation service, resulting in improved fairness, distribution outcomes, algorithmic awareness, and identification of decision-making inconsistencies. In a similar vein, [40] explore the permissibility of actions that save lives while causing harm to others, proposing a computational model using subjective utilities to capture moral judgments across diverse scenarios, extending beyond typical life-and-death dilemmas and suggesting integration with causal theory.

In this work, we focus our attention on Conditional Preference networks (CP-net), a graphical model for compactly representing conditional and qualitative preferences [20]. 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. Given a CP-net, one can address optimality questions, that look for the most preferred decision, or also dominance questions, that ask for comparing the preference of two decisions. Many algorithms to respond to such questions, for several versions of CP-nets, have been defined, and their computational complexity has been thoroughly studied.

CP-nets have been extensively used in preference reasoning, preference learning, and social choice literature as a formalism for modeling and reasoning with qualitative preferences [51, 58, 59]. They have been used to compose web services [60] and to support other decision aid systems [61]. CP-nets are a particularly attractive formalism as they provide an explicit model of the dependencies between features of a decision: for instance, in the example of the two meals, the wine feature depends on the main dish feature. This is important for AI decision-making as, for instance, the Engineering and Physical Science Research Council (EPSRC) Principles of Robotics dictates the implementation of transparency in robotic systems specifically by providing a mechanism to "expose the decision-making of the robot [45]." Hence, having a formal, explicit decision model such as a CP-net can provide this transparency.

[2] Note that despite using human preferences as a core concept, preference-based models are not limited to capturing consequentialist or utilitarian models of moral cognition. As we illustrate below (§2.4 and 4), our formalism captures utilitarian as well as non-utilitarian elements of morality, such as agreement-based factors.

2.3 Value alignment and dual process models

The idea of teaching machines right from wrong has become an important research topic in both AI [62] and related fields [11]. This challenge has been pursued using a range of computer science approaches, including taking sequences of actions in a reactive environment [63] and teaching agents how to respond in specific environments [64]. Many of these projects address what is called the *value-alignment* problem [65], that is, the problem of building machines that behave according to values aligned with human ones [12, 66–69]. This aligns with the challenge of instilling morality into AI systems addressed by [70], who introduce Delphi, an experimental framework based on deep neural networks trained to reason about ethical judgments. This reveals the potential and limitations of machine ethics, emphasizing the necessity of explicit moral instruction and exploring alignment with ethical theories. We continue in this stream of research by proposing a novel and extensible formalism for modeling and reasoning with preferences over complex judgments.

Many human decision-making processes can be characterized by dual process models of cognition, often known as the "thinking fast and slow" approach [24]. Dual process theories describe the mind as composed of two broad reasoning approaches: (1) thinking fast, or System 1, which relies on predefined heuristics or rules, and (2) thinking slow, or System 2, which is more deliberate. Each of these systems has its merits. System 1 thinking is efficient, requiring only limited computational resources and contextual information. When System 1 heuristics are deployed in the environments they were intended for, they often produce good-enough decisions; though they sometimes can lead to sub-optimal choices when deployed in edge cases. System 2 decision-processes, on the other hand, are cognitively intensive and allow the decision-maker to flexibly integrate information from disparate sources, leading to decisions that can be carefully tailored to the case at hand. These decision-processes can be used in combination in a way that optimizes payoffs given the limitations of the computational resources available [71–73].

Dual process models have also been used to understand human moral judgment [26, 27, 74, 75]. These models draw their inspiration from two of the major branches of moral philosophy. The rule-based System 1 is inspired by *deontological* theories of moral philosophy, which focus on constraints on actions. The deliberative System 2 is typically associated with *consequentialist* theories, which focus on maximizing the utility of outcomes. System 1 applies a hard and fast rule and does not consider the subtleties or complexities of the current scenario. In contrast, System 2 acts more like a utilitarian reasoner, it attempts to quantify utility losses and gains, which enable it to make a judgment. For example, when Bob refrains from stealing something he cannot afford from a store even though he really wants it, he may be following a System 1 moral rule: no stealing. When Susan decides to donate $5 to buy mosquito nets rather than chocolate bars for children in crisis, she may be using System 2 reasoning and comparing the values of the outcomes of the candidate's actions.

2.4 Contractualism and universalization

Theories of moral psychology have successfully drawn on ideas from two major frameworks in moral philosophy—deontology and consequentialism—to contribute to our understanding of the moral mind. Curiously, the central idea of another family of philosophical views—*contractualism*—has been virtually absent from theories of moral

psychology. Contractualist views ask us to consider what agreements could be adopted that would lead to mutual benefit [76–79]. Recent work has begun to show that, despite being neglected for so long, contractualist mechanisms play a critical role in the moral mind [36, 80–83].

The specific agreement-based process we focus on in this paper is *universalization*—a psychological mechanism akin to Kant's Categorical Imperative [84]—which is a way of making a moral decision by asking "what if everyone felt at liberty to do that?" [36]. When people universalize, they imagine a hypothetical world where everyone is allowed to act in a certain way. If things go well in that hypothetical world, the action in question should be allowed. If things go badly, the action should be prohibited. This boils down to asking if a person is taking a special privilege for themselves that they couldn't grant to everyone. If everyone could feel free to do the action, then presumably *mutual agreement* could be reached that the action is permitted. Universalization is likely to be a System 2 process because running the universalization computation is resource-intensive and requires a lot of information about the particular decision-making context [36].

2.5 Extending the dual process model of morality

We propose an extension of standard dual-process models of morality [85]. Not only do humans have two systems that they can use to make moral judgments, but the two systems flexibly interact. Humans sometimes use outcome-based and agreement-based (System 2) processes to figure out when a (System 1) rule can be overridden [85].

This proposal is intuitively plausible. After all, sometimes people decide that it is morally acceptable to break previously established rules because doing so would bring about a better *outcome* than following the rule [26, 86]. For example, in general, there is a rule against non-consensual harmful contact, i.e., no hitting. But if pushing someone out of the way of a speeding train could save their life, then it is morally permissible to do so [87]. Yet in other cases, people may decide that it is morally acceptable to break a rule because the person whom the rule protects would *agree* to have the rule violated [80, 88]. There is a rule against taking things that don't belong to you (or: no stealing), but it may be morally acceptable to take coffee grounds from your co-worker's desk if you know they would consent to it, given that they are likely to be caught in a similar situation in the future and want to take some coffee from you. Following this pattern, we further hypothesize that universalization is another System 2 process that will help people determine when it is OK to break a rule.

In this work, we focus on the rule about waiting in line: "no cutting". Scenarios involving waiting in line provide an interesting test-case to study how people make moral judgments,[3] Lines are seemingly adjudicated by a simple rule that everyone can articulate. However, people's judgments about when it is permissible to cut in line are (at least somewhat) consistent and replicable, even in completely novel contexts that

[3] In this paper, we sometimes refer to rules that guide these cases as "socio-moral" and the judgments people render of them "socio-moral judgments." This is because there is no widely agreed-upon definition of what counts as a moral rule as opposed to a socially conventional one [89]. In fact, recent research shows that people from different cultures have different understandings of what counts as a moral issue to begin with [90]. Moreover, there is no consensus as to whether there are separate psychological processes for judging moral rules verses social ones [89]. For our purposes, it is not critical whether the rule about standing in line is moral *per se* but rather that it exhibits certain interesting characteristics.

participants have never seen before [91, 92]. This suggests that participants are not simply "memorizing" the exceptions to the rule, but instead using a generative psychological mechanism to make judgments in novel cases [85, 91, 92]. This generative mechanism is what our computational formalism is designed to describe and be aligned with.

3 Formal background: CP-nets

Conditional Preference networks (CP-nets) are a graphical model for compactly representing conditional and qualitative preferences. 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 preference reasoning, preference learning, and social choice literature as a formalism for working with qualitative preferences [51, 58, 59]. CP-nets have even been used to compose web services [60] and other decision aid systems [61]. While there are many formalisms to choose from when modeling preferences, we focus on CP-nets as they are graphical and intuitive.

We borrow from [20] the definition of a CP-net:

Definition 1 A CP-net over variables $V = \{X_1, ..., X_n\}$ is a directed graph G over $X_1, \ldots, X_n$ whose nodes are associated with conditional preference tables $CPT(X_i)$ for each $X_i \in V$. Each conditional preference table $CPT(X_i)$ associates a total order $\succ_u^i$ with each instantiation u of X_i's parents $Pa(X_i) = U$ in the directed graph.

A CP-net has a set of features (also called variables) $V = \{X_1, \ldots, X_n\}$, each with a finite domain $D(X_1), \ldots, D(X_n)$. 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 directed *dependency graph* G 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 the 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 takes the form of a total or partial order over $D(X_i)$. This is formally done via the notion of cp-statement. Given a variable X_i with domain $D(X_i) = \{a_1, \ldots, a_m\}$ and parent variables $Pa(X_i) = \{Y_1, \ldots, Y_n\}$, a cp-statement for X_i has the form: $Y_1 = v_1, Y_2 = v_2, \ldots, Y_n = v_n : X_i = a_1 \succ \ldots \succ X_i = a_m$, where, for each $Y_j \in Pa(X_i)$, $v_j \in D(Y_j)$. The set of cp-statements for a variable X_i is called the cp-table (CPT) for X_i. A full example is given as Example 1

Among the several generalizations of CP-nets, in this work, we focus on one that models cases where individuals have indifference over the values of some features or where they do not specify some preference information [93]. A cp-statement $z : a_i \approx a_j$ means that the person is indifferent between a_i and a_j given the assignment z to the parents variables, i.e., $a_i \succeq a_j$ and $a_j \succeq a_i$. A lack of information over one of the values of a variable is modeled with empty cp-statements for that value.

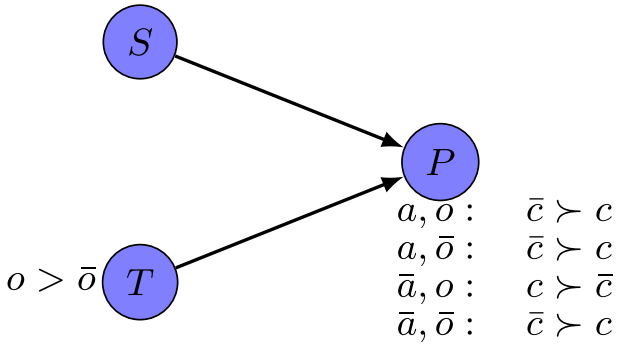

Fig. 1 The CP-net representing John's preferences described in Example 1. Variable S denotes the scenario and has the values a for "at the airport", and $\overline{a}$ for "not at the airport". Values in the domain of this variable are incomparable because no cp-statements appear in the CP-table of this variable, i.e., we assume John is indifferent between these values; variable T represents time, with values o for "on time", and $\overline{o}$ for "late"; variable P represents the preference over letting people cut the line, with values c is for "ok to cut the line", and $\overline{c}$ is for "not ok to cut the line"

Example 1 This example provides a CP-net for a scenario that is related to the data we have collected (discussed later in the paper), that has to do with people expressing a moral judgment over the action of cutting the line.

We consider a person, say John, who is the first in line at the airport security. A person approaches him and asks to cut the line because her flight is going to leave very soon. Because of security considerations, John always prefers not to let anyone cut the line at the airport. However, he has different preferences when he is not at the airport: he lets a person cut the line if he is on time, while he prefers not to let someone cut the line if he is late.

The CP-net in Fig. 1, with features S, T, and P, represents John's preferences as described in Example 1: variable S denotes the scenario and has values a for "at the airport", and $\overline{a}$ for "not at the airport", variable T represents time, with values o for "on time" and $\overline{o}$ for "late", and variable P models his preference over the request and has values for yes (c) and no ($\overline{c}$). In this CP-net, S and T are parent variables to P. The first cp-statements for P states that, when he is at the airport ($S = a$) and he is on time ($T = o$), he prefers to not let people cut the line ($\overline{c} \succ c$). A similar interpretation can be made for the other three cp-statements for P. For variable T, we assume that John prefers to be on time rather than late, so we include the cp-statement $o \succ \overline{o}$. On the other hand, for the scenario variable S, we do not assume any preference because values in the domain of this variable are incomparable due to the fact that no cp-statements appear in the CP-table of this variable.

The semantics of a CP-net depends on the notion of a *worsening flip*, which is a change in the value of a variable to a less preferred value according to the cp-statement for that variable. One outcome (that is, an assignment of values to all the variables) 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 of dominance induces a preorder (i.e. a binary relation which is reflexive and transitive) over the outcomes. Indifference induces a loop between pairs of outcomes while the lack of information induces incomparability in the preorder (i.e. given two outcomes o, p, if neither $o \succeq p$ nor $p \succeq o$ are valid, then we say that o and p are incomparable, denoted with $o \bowtie p$.). This incomparability makes the preference graph disconnected. In particular, in the induced preference graph there is a connected

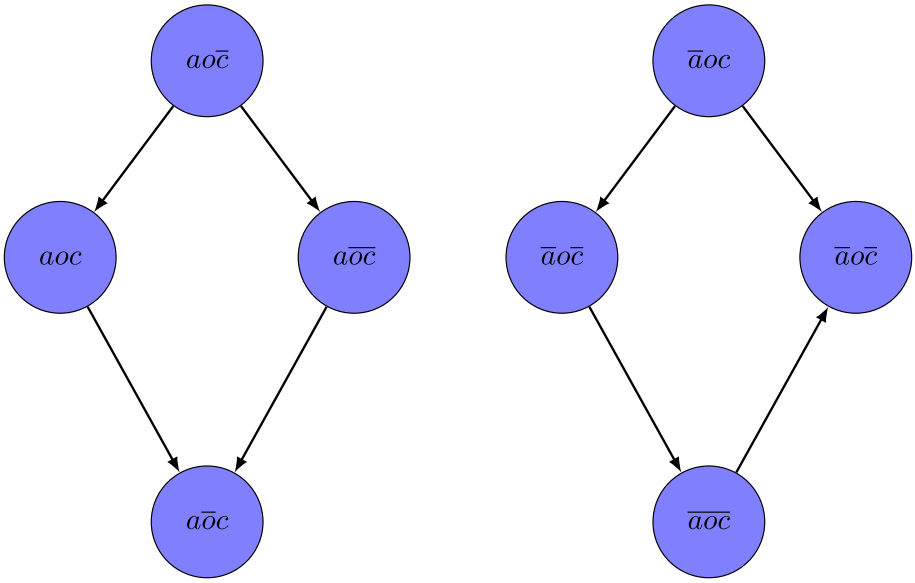

Fig. 2 The preorder induced by the CP-net in Fig. 1. The two components are due to the incomparability on the domain of the variable S

component for each combination of values of the variables with missing cp-statements. For instance, Fig. 2 gives the full induced preference order for the CP-net shown in Fig. 1. The component on the left side describes preferences over the airport scenario ($S = a$), while the component on the right side describes John's preferences when not at the airport ($S = \overline{a}$). Clearly, outcomes in different components cannot be compared because they describe preferences over different scenarios. In this, incomparability is a useful tool that allows us to model different scenarios.

While CP-nets are usually a compact way to express preferences, their induced order can be exponentially larger. This is why it is important, for computational sake, to be able to reason on CP-nets and not on their induced graphs. Two important questions about CP-nets are related to optimality and dominance [58]. Finding the optimal outcome of a CP-net is NP-hard [20] in general but can be found in polynomial time for acyclic CP-nets, by assigning the most preferred value (according to the CP–tables) for each variable in the order given by the dependencies. Indeed, acyclic CP-nets induce a lattice over the outcomes as depicted in Fig. 2. The induced preference ordering, Fig. 2, can be exponentially larger than the CP-net itself as shown in Fig. 1. On the other hand, checking the dominance between two outcomes is a computationally difficult problem [94].

4 SEP-nets: scenarios, evaluation, and preferences for modeling 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 evaluation 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 when such reasoning takes place outside the CP-net formalism.

Given this limitation, there have been calls to extend this model into reasoning and preference systems in computer science in a principled and exact way [52]. We therefore propose an extension to the CP-net formalism to be able to capture the psychological mechanisms at play in our humans' moral judgments. Following our discussion of human moral decision-making in Sect. 2, we hypothesize that humans employ a combination of System 1 (rule-based) and System 2 (consequentialist and contractualist) thinking. This insight leads us to model a two-modality reasoning process in an extension of the CP-net formalism so that we can embed it within a machine, to allow the machine to reason both fast and slow about ethical principles [18, 95]. Our motivation is to extend the semantics of CP-nets in order to model both the snap judgments that do not take into account the particularities of the scenario, System 1 thinking, as well as provide the ability to reason about these details, using System 2 thinking, if necessary [15].

Informally, our proposed generalization of the CP-net formalism is called SEP-nets (Scenarios, Evaluation, and Preferences networks), to handle variables associated with the context. We propose extending the formalism consisting of:

- a set S of *scenario variables* (we call SVs in the text) to define a decision-making context over which there is no preference to be stated;
- a set E of *evaluation variables* (we call EVs in the text) to model the evaluation process that takes place in the subjects' minds while reasoning over the given context;
- to decide their preference over the *preference variables (set P)* (we call PVs in the text) that are already modeled in CP-nets.

SEP-nets allow us to have a compact framework to express and reason about knowledge on moral preferences. Notice that when SEP-nets include only preference features, they collapse to being standard CP-nets. SEP-nets are built on a generalization of CP-nets proposed by [93] allowing for incompatibility in the set $Eval(X)$, which is not captured in the cp-tables of traditional CP-nets. For SEP-nets, we use incomparability to model components in the preference graph whose nodes cannot be compared. For instance, outcomes that depend on specific features of the scenario variables, i.e., the context of the decision, cannot be compared as these are not under the control of the decision maker. This change, with the addition of the evaluation functions *ef* for real-valued judgements mean that the SEP-net model is a strict generalization of the classic CP-net model.

Definition 2 An SEP-net consists of

- A set of features (or variables) $V = S \cup E \cup P$. Given a variable X, the domain of X is a finite set if $X \in S$ or $X \in P$. In particular, if $X \in E$ then the domain of X is $< EF, Eval(X) >$, where EF is a set of evaluation functions and $Eval(X)$ is a set of evaluation values. Each variable can be only in one of three sets, that is, $S \cap E = \emptyset$, $S \cap P = \emptyset$, and $P \cap E = \emptyset$.
- As in a standard CP-net, each variable X has a set of parent variables $Pa(X)$, on which it depends on. However, $Pa(X) = \emptyset$ if $X \in S$; $Pa(X) \subseteq S \cup E$ if $X \in E$, and $Pa(X) \subseteq S \cup E \cup P$ if $X \in P$. This models a three-level acyclic structure, where scenario variables are independent, evaluation variables can depend only on scenario

variables or other evaluation variables, and preference variables can depend on any variable.

- If $X \in E$, then given an assignment u to $Pa(X)$ and a value $j \in Eval(X)$ an evaluation function $ef_u(j) \in EF$ identifies a single value in [0, 1]. If $X \in P$, a standard CP-table states the preferences over the domain of X: for each combination of values in $Pa(X)$, the table provides a total order of $Dom(X)$. No CP-table or evaluation function is associated with scenario variables.

Preference variables may depend directly on either all or a subset of the scenario variables. Moreover, they may depend on certain values of a scenario variable, but not others. This direct dependency between preferences and scenarios models a sort of System 1 approach, where people make a moral judgment (or any other preference decision) by just looking at the situation at hand and without performing any sophisticated reasoning. On the other hand, when preference variables depend on evaluation variables, which in turn depend on scenario variables, we model a sort of System 2 approach, where people consider a scenario and a preference question, and make an estimate of the consequences of the various options before finalizing their decision or judgment.

When there are no scenario nor evaluation variables, but just preference variables, an SEP-net is the same as a standard CP-net, so its semantics are defined as usual for CP-nets [20]. When instead we have a full SEP-net, its semantics is an order over all the SEP-outcomes, where a SEP-outcome includes not only assignments to the PVs but also all the SVs and EVs. For the sake of readability, in the following, we denote with lowercase letters an assignment to a variable (i.e., $X_i = x_i$ or $X_i = x_i'$). Given two outcomes $o_v = [s_1, \dots, s_n, e_1, \dots, e_m, p_1, \dots, p_k]$ and $o'_v = [s'_1, \dots, s'_n, e'_1, \dots, e'_m, p'_1, \dots, p'_k]$, we have $o'_v \succ o_v$ if

- $[s_1 \dots, s_n] = [s'_1 \dots, s'_n]$. This means that we are considering the same scenario.
- $[e_1 \dots, e_m] = [e'_1 \dots, e'_m]$. That is, evaluation variables are set to their estimates as given by their evaluation functions, based on a vector of evaluation values $v = [v_1, \dots, v_m]$, with $v_i \in Eval(E_i)$.
- $[p'_1 \dots, p'_k] \succ_p [p_1 \dots, p_k]$ in the order $\succ_p$ induced by the CP-net which is obtained by just considering the preference variables (in P) and the dependencies among them.

So, outcomes that differ in the scenario are not connected in the order induced by a SEP-net. Moreover, any outcome with the evaluation variables set to values that are different from the value given by their evaluation function are not connected to any other outcome. If the dependencies among preference variables define an acyclic graph, the result is a set of partial orders, one for each scenario. Each of such partial orders has the same shape as the induced order of the CP-net obtained by the given SEP-net by setting the scenario variables to any value in their domain and the evaluation variables to their estimated value. Given an SEP-net and its induced set of partial orders, the optimal outcomes are the top elements in the induced partial orders, thus one for each scenario, and can be efficiently computed by 1) choosing any scenario, 2) setting the evaluation variables to their estimate value, given the chosen scenario, and 3) taking the most preferred values for the preference variables.

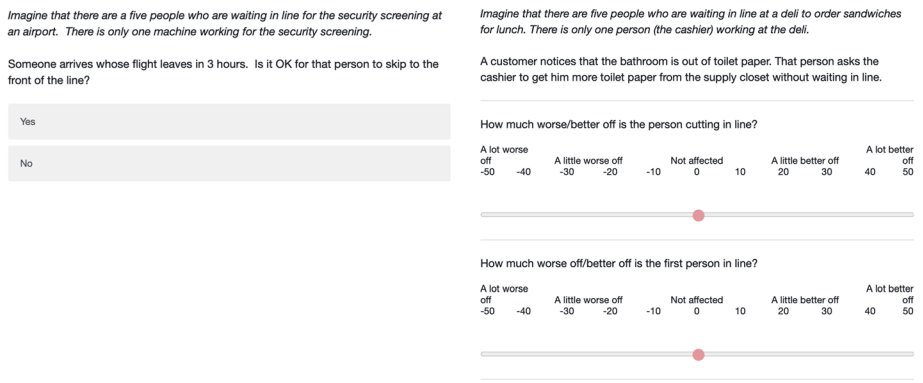

Fig. 3 Screenshot of the experimental set-up. Left Panel: Moral acceptability judgments. Right Panel: Evaluation questions

Table 1 Variables that went in to constructing each scenario

Variable	Scenario description
Reason	The particular reason for cutting the line, see Table 3
Location	The particular location of the scenario: *Deli*, *Airport*, or *Bathroom*
Main service	Whether the individual cutting the line had the goal of accessing the main service

5 Empirical evaluation

To test our model, we ran an experiment on Amazon MTurk. Informed consent was given by all participants and this study was approved by the Massachusetts Institute of Technology Institutional Review Board.[4] 407 subjects participated in the study in 2020. Following attention checks, the data from 301 subjects was retained for analysis. No demographic data was taken from participants, but the average demographic information for MTURK participants is as follows [96]: Gender, 55 % Female; Age, 20% born after 1990, 60% born after 1980, and 80% born after 1970; Median household income, $47K/year.

Subjects were randomly assigned to one of three story contexts, in which subjects were asked to imagine that they were standing in line as a *deli* (12 scenarios), for a single-occupancy *bathroom* (7 scenarios), or at an *airport* security screening (6 scenarios). (Refer to Table 1 for details.) These three contexts were selected because they represent mundane, everyday situations that call for moral judgment, a kind of "mundane realism" [97]. Judgments in these cases, therefore, are likely to represent our participants' commonly deployed moral reasoning capacity. We expect patterns of judgments in these cases to generalize to other cases of waiting in line – and ideally other cases of rule-breaking.

[4] The Committee on the Use of Humans as Experimental Subjects. The review process ensures compliance with university-mandated ethical guidelines for research conducted with human subjects. See couhes.mit.edu for details of the review process.

Table 2 Questions that each respondent was asked in order to evaluate each possible scenario

Variable	Prompt
Global Welfare	Think about the well-being of all the people in line combined. How are they affected by the person cutting in line?
First Person Welfare	How much worse off/better off is the first person in line?
Middle Person Welfare	How much worse off/better off is a person standing in the middle of the line?
Last Person Welfare	How much worse off/better off is the last person in line?
Line Cutter Welfare	How much worse off/better off is the person that cut in line?
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?
Likelihood	On any given day, how likely is it that this scenario going to happen?
Judgement	Is it acceptable to cut the line? (yes or no).

These contexts present the opportunity for someone to want to cut in line for a diverse range of reasons. We developed cases that manipulated (1) the amount of time by which the person cutting would delay the line, (2) the benefit that the person cutting would accrue by cutting, (3) the benefit that the people waiting would accrue by this person cutting, (4) the likelihood this particular scenario would happen at all. We hypothesized that if subjects were using agreement-based or utility-based reasoning to make their judgments, then these manipulations would have systematic impacts on moral acceptability judgments.

After reading each vignette, subjects were asked about the acceptability of cutting in line in that scenario: *"Is it OK for that person to ask the cashier for a new spoon without waiting in line?"* Subjects were allowed to answer "yes" or "no". Complete descriptions of the scenarios are given in Table 3, and Fig. 3 gives a screenshot of the experimental set-up.

As a way of evaluating the role of System 2 reasoning in producing these moral judgments, we ask subjects to evaluate each scenario on a series of utility and agreement-based measures based on our experimental manipulations. To evaluate the role of *System 2 outcome-based thinking*, Subjects were asked to estimate the utility consequences of the person cutting in line in each scenario, e.g. how long the cutter would delay the line, the benefit to the cutter, the detriment to the line. To evaluate the role of *System 2 agreement-based thinking*, we ask subjects what would happen, overall, if this type of line-cutting were always allowed, a proxy for whether everyone would agree to allow this person to cut [36]. Subjects responded on a scale of -50 (a lot worse off), 0 (not affected), +50 (a lot better off); likelihood was judged between 0 (not likely) and 100 (very likely). For a full list of these "evaluation questions" see Table 2.

Half the subjects were shown the evaluation questions for all the scenarios followed by the permissibility questions for all the scenarios; the other half of the subjects received the blocks of questions in the reverse order. This was designed to test and eliminate if necessary, the effects of evaluation then judgment versus judgment then evaluation.

Finally, we coded each vignette for whether or not the person attempting to cut had the goal of accessing the main service the line was providing. This allowed us to check whether a slight addendum to the rule about waiting in line could explain our findings. That is, rather than the rule simply being "no cutting in line", perhaps the rule participants use is "no cutting if you are interested in accessing the main service." The main service for the deli line was the sale of an item, for the airport scenario it was security screening, and

Table 3 Description of all 25 scenarios used in our experiments

Label	Scenario Description	Main Func.	Already Waited
Deli Line			
Spoon	A customer who is eating soup at the deli dropped his spoon on the floor and needs another one.	False	True
Water	A customer who is eating lunch at the deli wants a refill on tap water.	False	True
Soda	A customer who is eating lunch at the deli wants to buy another soda.	True	True
Catering Order	A customer wants to ask a series of questions about a catering order that he will pick up next week.	False	False
Fasted	A customer walks in who has just finished fasting for 24 h for a colonoscopy and is extremely hungry.	True	False
Diabetic	A customer walks in who is diabetic and urgently needs sugar.	True	False
Oven Repair	The oven-repair technician shows up and needs to ask the cashier a series of questions about the oven so he can fix it.	False	False
Soap	A customer uses the last of the hand soap in the bathroom.	False	False
Toilet Paper	A customer notices that the bathroom is out of toilet paper.	False	False
Spouse	A customer walks in who is married to a customer who is currently placing an order with the cashier.	True	False
Father	The father of the family is currently placing an order with the cashier.	True	False
Sandwich	A customer walks in who wants to order a sandwich.	True	False
Bathroom Line			
Wash Hands	Someone at the back of the line just needs to wash their hands.	False	False
Cleaner	Someone arrives who needs to clean the bathroom.	False	False
Vomit	Someone at the back of the line needs to throw up immediately.	True	False
Get Jacket	Someone at the back of the line thinks they forgot their jacket in the bathroom.	False	True
Friend	Someone at the back of the line is a friend of someone at the front of the line.	True	False
Aid	Someone at the back of the line is an aid to an elderly person at the front of the line.	True	False
Use Bathroom	Someone at the back of the line needs to use the bathroom.	True	False
Airport Security			
Departure in 20min	The flight leaves in 20 min.	True	False
Crying Baby	Standing with a baby who is crying very loudly.	True	False
Forgot Jacket	Forgot their jacket at the check-in counter.	False	True
Cafe Worker	Works at a cafe inside the airport.	False	False

Table 3 (continued)

Label	Scenario Description	Main Func.	Already Waited
Go to Bathroom	Has to leave the line to go to the bathroom.	True	True
Departure in 3 h	The flight leaves in 3 h.	True	False

"Main Func." refers to whether or not the person cutting is attempting to cut to receive the main function or service that the line is providing (i.e., to purchase an item in the deli cases, to get through security in the airport cases, and to use the toilet in the bathroom cases). "Already Waited" refers to whether the cutter already waited in line once before attempting to cut

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, one might view the main service of the deli line as receiving anything from the cashier, rather than purchasing something. Our characterizations of the main function of the line are rough approximations meant to describe one view that seems commonly held. This coding can be found in Table 3, see column titled "Main Func.". We also coded each scenario for whether or not the person attempting to cut had already waited.

At the end of the survey, participants were given an attention check as follows: "Finally, we are interested in learning some facts about you to see how our survey respondents answer questions differently from each other. This is an attention check. If you are reading this, please do not answer this question (do not check any of the boxes). Instead, in the box below labeled 'Other' please write 'I am paying attention'. Thanks very much!" This was followed with a list of levels of education and a free-response box labeled "Other." Participants who checked any box or failed to write "I am paying attention" in the appropriate place were screened out of the study.

5.1 Data analysis

We perform an extensive analysis of the collected data to uncover possible relationships among the variables. We will then leverage these insights in testing the applicability of the SEP-net model in the next section.

Figure 4 shows overall moral permissibility data, i.e., average OK or not-OK judgments, for each of the scenarios. The most notable feature of this data is that moral permissibility is graded; cutting the line is endorsed probabilistically by our subjects rather than in an all-or-none fashion. This is the first hint that our subjects are not using a simple rule to figure out when it is permissible to cut in line. A rule like "don't cut" for instance, would produce unanimously low permissibility for all cases. A slightly more sophisticated version of this rule-based approach would be that subjects use the rule "don't cut" but realize that the rule is not operative in certain scenarios. This would yield a slightly more complex rule such as "cutting is allowed only when you are not requesting the main service." This would yield a binary all-or-none pattern of results, with some instances of cutting being permissible (i.e. the red bars in Fig. 4) and some not (i.e. the blue bars). Instead, it seems that a more sophisticated understanding of the computations behind subjects' moral judgments is required. The aim of our model is to be able to predict our participants' graded permissibility judgments.

We checked if the order of question presentation (whether the subjects were asked the evaluation questions or moral judgments first) impacted moral judgments. To test this, we ran a Wilcoxon signed-rank test[5] against the null hypothesis that *changing the*

[5] The Wilcoxon signed-rank test [98] is a non-parametric test for comparing paired data samples from the evaluations of individuals. A Non-parametric test means we do not make assumptions about an underlying distribution of the data, e.g., we do not assume our data follows a normal distribution. The Wilcoxon signed rank test assumes that there is information in the magnitudes and signs of the differences between paired observations. It is considered the non-parametric equivalent of the paired student's *t*-test. In particular, the signed-rank test can be used as an alternative to the *t*-test when the population data does not follow a normal distribution. It is used to test the null hypothesis that two related paired samples come from the same distribution.

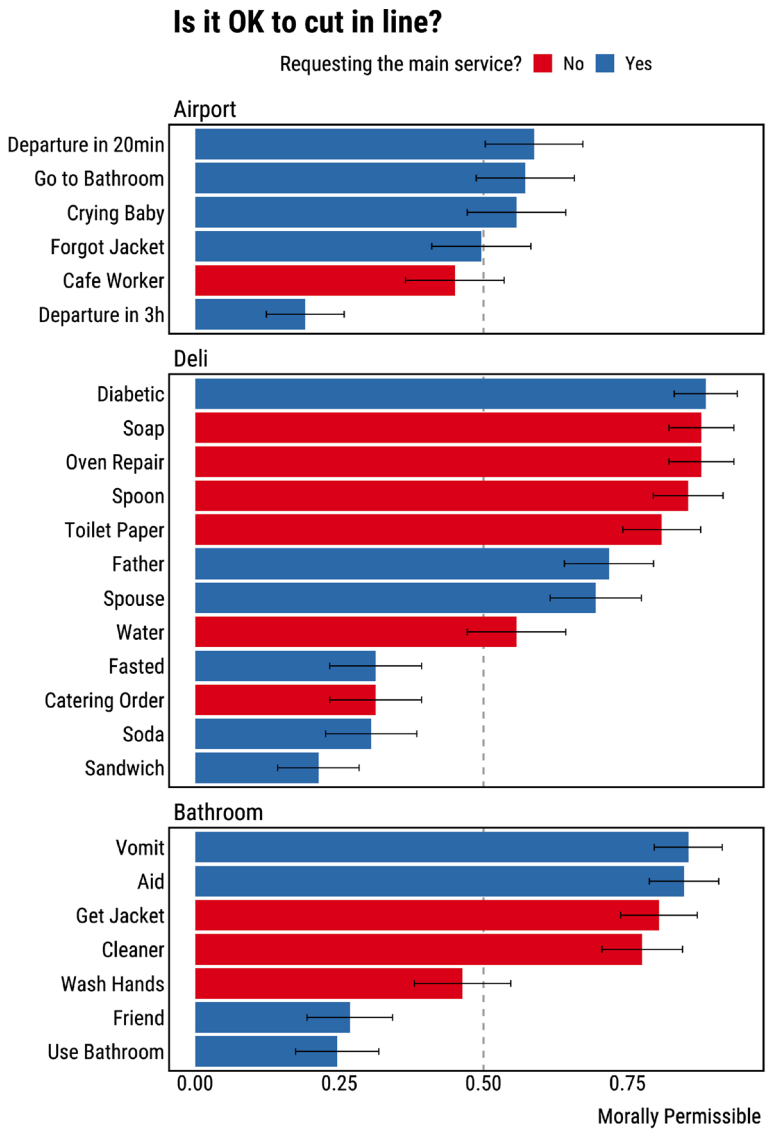

Fig. 4 Moral judgments about cutting in line in each of the scenarios. Color indicates if the person cutting in line is requesting the main service or not (blue for Yes, red for No). Error bars are 95% confidence intervals. As we can see, a simple rule, such as "it is ok to cut if you are not requesting the main service" is not sufficient to explain variation

order of evaluation and judgment does not influence the judgment value. We ran this for all 25 scenarios described in Table 3, and the full set of p-values is available in the supplemental material. The test did not reject the null hypothesis for any of the 25 scenarios. Hence, under the conditions of this study, asking individuals to think closely about the evaluation questions did not cause them to change their judgments. We conjecture that this happens because people already went through the internal evaluation

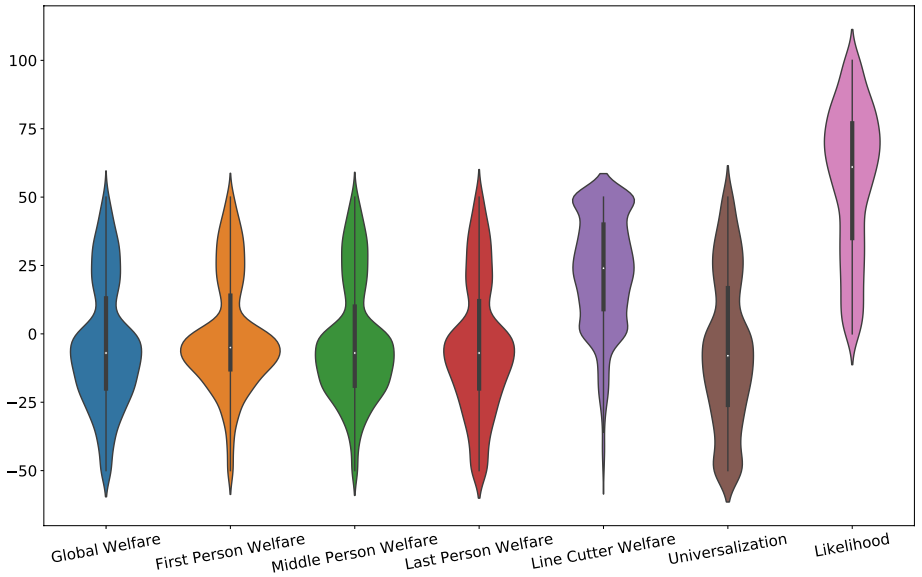

Fig. 5 Violin-plots depicting the distribution of responses to the evaluation questions. The width of the violin plot at any point is the number of responses with that value

process when they made a decision. Thus asking for it before or after does not influence the decision they made. We therefore pooled subjects in both order conditions for the remaining analyses.

Figure 5 depicts the distribution of responses by all participants for the evaluation questions. We conjectured that the different evaluation variables might draw out that individuals found various elements of the scenarios more or less important, allowing for the emergence of a better predictor by using multiple variables. For instance, the well-being of the first person could matter most because she has the most to lose if someone cuts in line, i.e., her change in wait time is proportionally greatest. On the other hand, it might emerge that the last person is most important because waiting additional time might actually make it no longer worthwhile to wait at all. Furthermore, *Global Welfare* or *Middle Person Welfare* might be the best because they provide an aggregate estimate or average estimate. As it turns out, people do not judge these metrics differently (see Fig. 5). Additionally, there is a fair amount of negative skew for the question *Line Cutter Welfare*, indicating that some participants felt that even if cutting the line was allowed, they were not receiving much benefit.

Finally, we wanted to see if there were any strong correlations between the various evaluation questions and the moral judgment. Figure 6 shows the cross-correlation between all responses from the subjects. As one might expect, the questions about the individuals in line are highly correlated, further indicating that most subjects respond to these questions with similar evaluations. In addition, as predicted, all evaluation variables are negatively correlated with the moral judgment variable, indicating that as there are more negative impacts of the action, the less likely it is judged permissible to cut in line. Universality has the strongest negative correlation with the moral judgment variable as well, indicating that participants seem to consider the question "what if everyone did this" (the System 2 contractualist method of reasoning) when deciding if it was OK to cut.

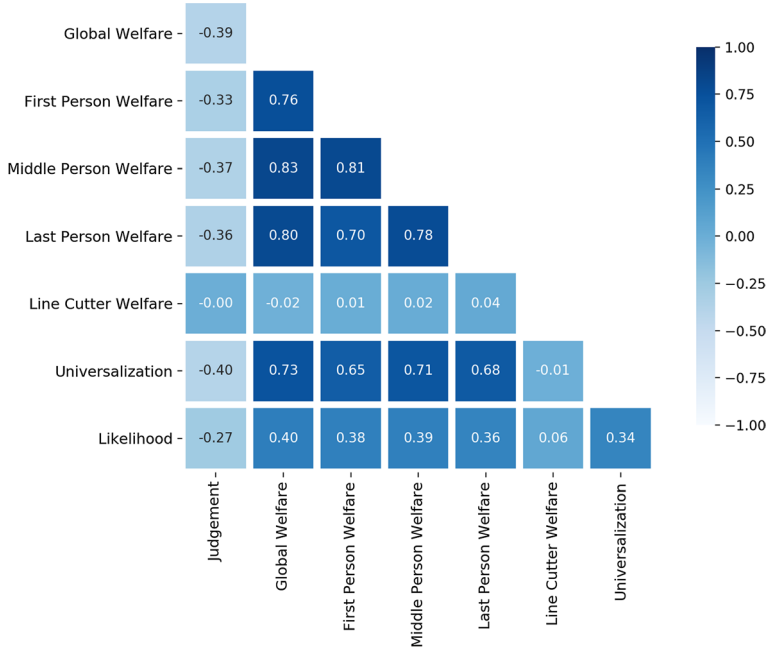

Fig. 6 Cross-correlation matrix for all scenario variables. The moral judgment is labeled as Judgement and is negatively correlated with all the evaluating metrics, indicating that as any measure gets worse, the moral permissibility goes down

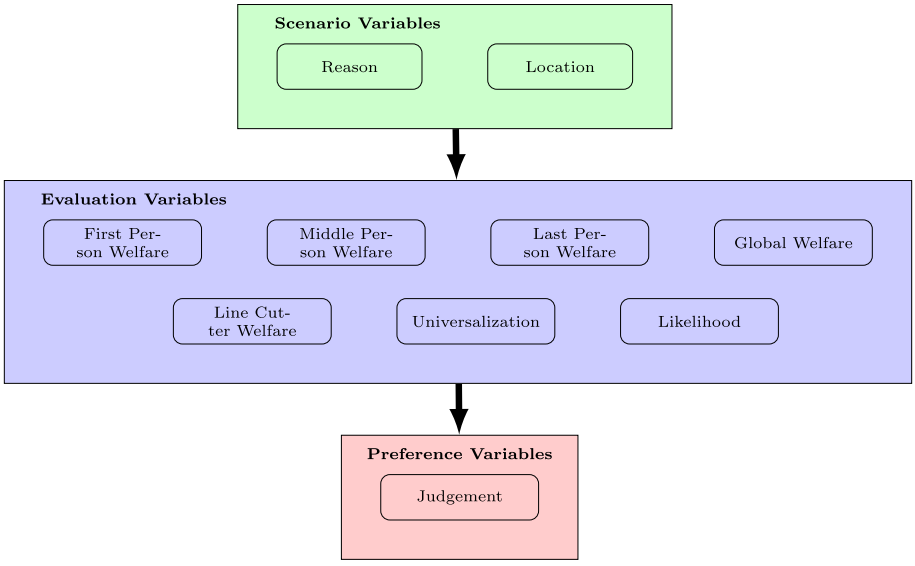

Fig. 7 Our model blends the notion of preferences with that of Scenario and Evaluation Variables. While individuals cannot set or have preferences over the Scenario Variables, they will possess their own subjective evaluations over the Evaluation Variables given a setting to the Scenario Variables. Given both the Scenario Variables and the Evaluation Variables, the agent can then decide on a preference over the single Preference Variable

5.2 Building an SEP-net from data

Leveraging the insights of our initial data analysis, we detail how we map the responses into an SEP-net that we will then use for predicting human responses to moral judgments in the next section. To start, a simple, fully connected version of an SEP-net, without accounting for the data, is depicted in Fig. 7. Working from our questionnaires we have the following variables:

1. *Scenario Variables (SVs).* 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, i.e., there is no incoming dependency arrow, meaning that this is part of the input to a decision-making AI system.
2. *Evaluation Variables (EVs).* 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 experimental details section. In our experiments, we have 8 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 are 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 evaluation phase in which they identify the evaluation variables and estimate a value for them.
3. *Preference Variables (PVs).* These are already included in a standard CP-net. In the setting under study, the agent expresses preference over a single value, that models whether or not, given both the values of the SVs and the values for the EVs, it is acceptable to cut the line, i.e., the moral judgment. The single preference variable depends on the evaluation variables. This again follows from the conjecture that a person needs to first perform a level of consequentialist or contractualist estimation in order to decide whether the rule, that states that a line cannot be cut, can be violated.

CP-nets and their variants, e.g., probabilistic CP-nets [23, 99], allow only for preference variables, and there is no option for creating a dependency between the preference variables to scenario and context variables. We rather envision a three-layer generalization where, as shown in Fig. 7, 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. Hence, we turn to look closer at our data in order to refine the fully connected SEP-net depicted in Fig. 7.

In order to build an SEP-net from our collected data, we must first identify dependencies between variables in our collected data. The variables of our SEP-net will be the same as those detailed above and the final result of our analysis can be seen in Fig. 8. The SVs

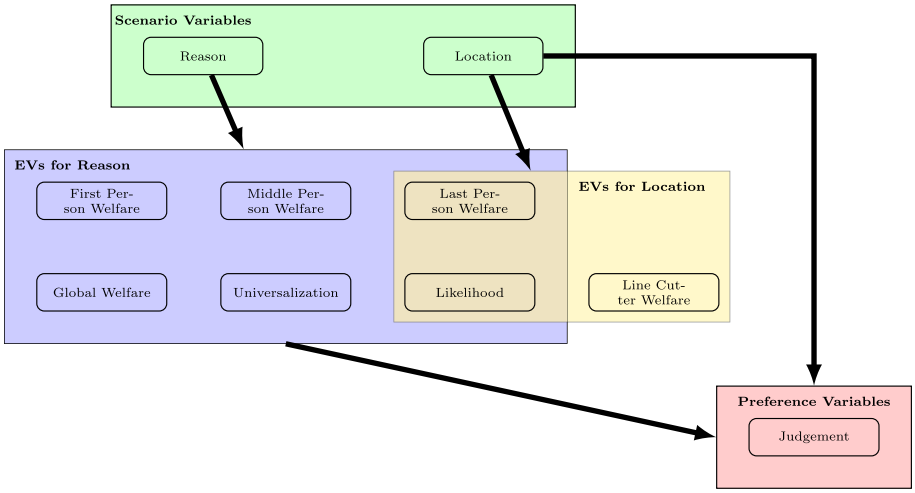

Fig. 8 The SEP-net corresponding to the data collected in our study. SVs influence the way individuals evaluate each scenario and make a decision. For the sake of readability, we group evaluation variables based on whether they depend on a particular SV in order to reduce the number of arrows. Given the SVs and the EVs, the agent can then decide on a preference over the single PV. Note that *Line Cutter Welfare* does not have any effect on the PVs, while the *Already Waited* variable is completely missing

describe all features of a scenario, and therefore refer to Reason and Location. Reason has a domain that includes all 25 reasons for cutting the line as listed in Table 3 and Location has domain $\{Airport, Bathroom, Deli\}$. The 7 EVs in Fig. 8 correspond to the evaluation questions asked to the subjects. All have the domain $[-50, 50]$, except Likelihood that has the domain [0, 100]. There is only one PV, corresponding to the moral judgment question. It has a binary domain (i.e., 0 “no cut", 1 otherwise).

To start, we investigate which SVs influence the way individuals respond to EVs. If we can find a relationship between these variables, then we can say that a subset of EVs depends on a subset of SVs, and use these relationships to build our model. We also check whether SVs influence the PV. To test for dependency, we run a set of Wilcoxon signed-rank tests to see whether the following four null hypotheses can be rejected:

1. NH1: Location does not affect the EVs;
2. NH2: Reason does not affect the EVs;
3. NH3: Location does not affect the PV;
4. NH4: Evaluation variables do not affect the PV.

5.2.1 NH1: Location does not affect all evaluation variables equally

Table 4 gives the p-values for our Wilcoxon signed rank test for each pair of location and evaluation variables. NH1 is only rejected in a few cases, specifically between Deli and Bathroom when evaluating the effect on the last person, cutter, and likelihood as well as between the Deli and Airport when evaluating the effect on the cutter and likelihood.

Table 4 p-values for the Wilcoxon signed-rank test against the null-hypothesis NH1: the location does not affect evaluation variables

Scenario	Deli-Bath	Deli-Airpt	Airpt-Bath
Global Welfare	0.2782	0.8954	0.3028
First Person Welfare	0.1779	0.0932	0.9696
Middle Person Welfare	0.1012	0.1390	0.3478
Last Person Welfare	**0.0064**	0.1444	0.2763
Line Cutter Welfare	**0.0069**	**0.0123**	0.3467
Universalization	0.4848	0.4356	0.1567
Likelihood	**0.0008**	**0.0138**	0.2430

The null-hypothesis is rejected if $p < 0.05$, these cases are reported in bold

Table 5 p-values for the Wilcoxon signed-rank test against the null-hypothesis NH3: location does not affect the preference variable

Scenario	p-value	Rejected
Deli-bath	8.759E−01	False
Deli-air	1.548E−08	True
Air-bath	2.662E−10	True

The null-hypothesis is rejected if $p < 0.05$

Hence we can say that *NH1 is only partially rejected, and that some Evaluation Variables are affected by the Location, specifically: last person, cutter, and likelihood*. We have depicted these results in Fig. 8 where we build an SEP net from our experimental data. It is interesting to notice that there are a set of variables that seem to be independent of location, e.g., Global Welfare is always important, but likelihood is dependent on location.

5.2.2 NH2: Reason does not affect the evaluation variables

For each pair of scenarios, we checked the Wilcoxon signed-rank test for each of the EVs; we omit the full $7 \times \binom{25}{2}$ pairs for readability (they are available in the appendix). For NH2 we can only reject the null hypothesis in some cases, however, *some evaluation variables are significantly affected for each assignment to the reason variable*. Hence, we can say that individuals evaluate the scenario differently based on why someone wants to cut the line, and that the reason can and does influence all evaluation variables, with the exception of the Line Cutter Welfare variable, as shown in Fig. 8.

dummy

5.2.3 NH3: Location does not affect the preference variable

We investigated whether or not the location had an effect on the PV. In order to test this, we selected four reasons for each location, since there were different numbers per location, aggregated these, and compared the mean responses to the moral judgment (PV). The complete results are depicted in Table 5. From this, we can reject NH3 for all pairs except Deli and Bathroom. *This indicates that in some cases location may be sufficient to evaluate*

the vignette and make a decision. This is represented by the arc between Location and PV in Fig. 8. It is interesting to note that participants seemed to evaluate the airport as being a significantly different location than the deli or the bathroom.

5.2.4 NH4: Evaluation variables do not affect the preference variable

In order to assess the influence of the EVs over the PV we need to place them all in a comparable range. To do this we compute quartiles of each evaluation variable to bucket them and then compare the effect of this group on the PV using the Wilcoxon signed-rank test. For each combination of SV and EV, we pair groups in order to perform the test. The full results are available in the appendix. *The results suggest that the EVs have some influence on the preference variable, except for the cutter variable.* Indicating that in general participants were not thinking of the benefit to the cutter, but only the cost to others. This is depicted in Fig. 8 as we have moved the cutter variable out of the box of the rest of the EVs, since is perceived differently.

Based on the findings we can construct a partial graph of the dependencies between variables, the resulting SEP-net corresponding to the data of our study is shown in Fig. 8. In the next section, we will put our SEP-net to the test as a prediction model for human moral judgments.

5.3 Prediction analysis

The analysis conducted in the previous section can be leveraged to automatically construct an SEP-net, provided suitable data. In this work, we use the SEP-net from data as a proof-of-concept system and test it using a prediction task, comparing its performance with that of several popular machine learning models. The aim of the task is to model a social behaviour in a specific scenario, i.e., given a location and a reason, and predict whether an individual would allow another individual to cut in line given her evaluations for each EV.

All models are trained on a 5-fold cross-validation. Meaning that the dataset has been split into 5 non-overlapping subsets. One subset is used for testing and the remaining four for training. The process is repeated 5 times, changing at each iteration the subset used for testing. In a binary prediction task, we evaluate models according to their performance as a combination of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) predictions. We employ the following standard metrics to evaluate the models:

Accuracy is the proportion of correct predictions out of all possible predictions, computed as: $\frac{TP+TN}{TP+TN+FP+FN}$.
Precision is the fraction of true positives among all positive predictions, computed as: $\frac{TP}{TP+FP}$.
Recall is the fraction of true positive predictions among all actual positives, computed as: $\frac{TP}{TP+FN}$.
F1 Score is the harmonic mean of precision and recall, computed as $\frac{2*TP}{2*TP+FP+FN}$.

Following the relationships discovered in the data analysis reported in Sect. 5.1, we modeled two SEP-nets. The two SEP-nets differ in the set of evaluation functions and evaluation values. The two SEP-nets have the same structure, as reported in Fig. 8. In particular, in one SEP-net we refer to as SEP-SVM, evaluation functions are modeled by letting each

Table 6 Quartile of the responses for each question. These quartiles were used to group the responses of individuals for each variable

Variable	0	1	2	3
Global Welfare	$[-50, -20]$	$(-20, 0]$	$(0, 10]$	$(18, 50]$
First Person Welfare	$[-50, -12]$	$(-12, -3]$	$(-3, 20]$	$(20, 50]$
Middle Person Welfare	$[-50, -18]$	$(-18, -5]$	$(-5, 17]$	$(17, 50]$
Last Person Welfare	$[-50, -20]$	$(-20, 0]$	$(0, 18]$	$(18, 50]$
Line Cutter Welfare	$[-50, 9]$	$(9, 25]$	$(25, 40]$	$(40, 50]$
Universalization	$[-50, -25]$	$(-25, -5]$	$(-5, 21]$	$(21, 50]$
Likelihood	$[0, 38]$	$(38, 63]$	$(63, 79]$	$(79, 100]$

one be a Support Vector Machine (SVM) [30]. SVMs are one of the most robust prediction methods for binary classification tasks and are commonly used for these tasks. In our model, given an evaluation variable, a location, and a reason, the correspondent evaluation function is trained on the subset of individuals' responses in the training set for that variable to predict the preference value. In SEP-SVM, the set of evaluation values coincides with the set of original values in the MT survey (i.e., values are in $[-50, 50]$ for all variables, except for Likelihood whose values are in $[0, 100]$).

In the other SEP-net, which we call SEP-Table, evaluation functions are modeled using the empirical distribution of the preference values in the training data, grouped by quartiles of evaluation values. This means, that given a variable, a location, and a reason, evaluation values in the training set are split into 4 groups, Table 6 gives the intervals for each evaluation variable. For each group, the function returns the preference value that is preferred by the highest number of subjects in the training set. In both SEP-nets, the CP-table of the preference variable is built by adopting a simple approach: each cp-statement reports the value that appears most often in the training set, given the assignment to the parent variables in the SEP-net. Notice that the assignment of the parents is given by the output of the selected evaluation functions except "line cutter welfare", and the value of the location variable. At training time we compute the number of individuals grouped by location and preference value. In building the CP-table, we use this value to break possible ties in the assignments of evaluation variables, i.e., in the case of ties, we choose the preference value that is preferred by the highest number of individuals for a given location. This models a simple majority rule on the evaluation variables with a tie-break rule.

Given a location, a reason, and a set of evaluation values of an individual, a SEP-net returns a prediction of whether the individual would allow someone to cut in line, consistent with the preferences in our collected data. We put SEP-nets to the test against a set of baselines leveraging popular machine learning models typically used for binary classification/prediction tasks. As baselines, we trained four models: an XGBoost [100, 101], an ensemble of 100 neural networks using VORACE [102], a Random Forest [103], and a single SVM. We did not perform any fine-tuning of hyperparameters but rather used their default values. The experiments were run on a MacBook Pro (13-inch, 2017), CPU 3.5 GHz Intel Core i7, RAM 8 GB 2133 MHz LPDDR3. The simulations are developed in Python 3.7. We adopted *RandomForestClassifier, SVM* by Scikit-learn, *XGBClassifier* by XGBoost.

The average performance and standard deviation across all five folds of our cross-validation, along with average training time, is reported in Table 7. We observe that the two SEP-nets perform very well overall, and slightly better than XGBoost which is considered

Table 7 Average performance on the test sets and the average training time of the different models in a 5-fold cross-validation, standard deviation in parentheses

Model	Accuracy	F1	Precision	Recall	Time (ms)
RandomForest	0.7651 (0.0069)	0.7119 (0.0190)	0.7402 (0.0121)	0.6859 (0.0276)	303 (24)
XGBoost	**0.7870 (0.0227)**	0.7307 (0.0417)	0.7822 (0.0325)	0.6868 (0.0556)	109 (4)
Vorace	0.7166 (0.0091)	0.6620 (0.0178)	0.6692 (0.0212)	0.6550 (0.0152)	181955 (10173)
SVM	0.7115 (0.0192)	0.6493 (0.0203)	0.6704 (0.0103)	0.6298 (0.0299)	6157 (454)
SEP-Table	**0.7870 (0.0261)**	**0.7329 (0.0367)**	0.7817 (0.0354)	**0.6906 (0.0438)**	**22 (1)**
SEP-SVM	0.7834 (0.0248)	0.7224 (0.0340)	**0.7926 (0.0458)**	0.6654 (0.0424)	259 (17)

Best performance in bold

state-of-the-art in many domains. In the case of SEP-Table, the training time is much smaller than the others since the model is mostly fit from the empirical data. It is interesting to notice that the single SVM performs much worse across all metrics than SEP-SVM. This result suggests that the specialization adopted inside the proposed model seems to be beneficial for modeling highly contextual decisions. The proposed SEP-nets have high accuracy with low standard deviation, showing that they are able to fit our data very well. From this model, one may conjecture that the computation of the preference value appears to be the output of a deliberation process that chooses a moral judgment based on the majority of more simple decisions.

6 Discussion

The analysis reported in Sect. 5.1 and the results reported in Sect. 5.3 achieve the two central goals of this paper. First, they provide evidence for our hypothesis about moral psychology, namely, that System 2 (outcome-based and agreement-based) reasoning is at play for our participants when deciding when to override rules. Second, they demonstrate that our novel formalism (SEP-nets) describes this process in a way that could be useful for enabling AI systems to understand the bounds of constraints.

There is a consensus in the moral psychology literature that rules matter for moral judgment [26, 27, 87, 104, 105]. However, what is less clear is how we readily figure out when rules can be overridden. In this paper, we proposed that a System 1 mechanism is often at play when we follow rules (as others have suggested before us, see especially [26]) and a (partially) contractualist System 2 mechanism is at play to figure out when to override rules (our original contribution). The data from our study supports this hypothesis. If our participants were using an entirely rule-based approach to make moral judgments about the scenarios we gave them, then they would have responded by 1) saying it was never permissible to cut in line (a strict rule-based approach), 2) saying it was only permissible to cut in line if you weren't requesting the main service (a slightly more nuanced version of the rule about standing in line), or 3) being responsive only to the location variable and not the evaluation variables. It is clear that participants were *sometimes* rule-based. This was revealed by the fact that the location variables sometimes have a direct influence on the preference variable (see Fig. 8). However, subjects also clearly used System 2 reasoning,

which was apparent by the fact that the evaluation variables (including both utility-based and agreement-based concerns) impacted the preference variable.

Constraints on AI systems can be useful to ensure that AI systems are abiding by the social and moral rules that guide the human world. However, constraints have their drawbacks. Humans navigate constraints quite flexibly, understanding when a seemingly fixed constraint should actually be overridden. SEP-nets present a formalism that is inspired by the way that human participants sometimes decide to override socio-moral rules. To generalize this formalism to other socio-moral rules, evaluation variables for those rules would have to be determined. In the simulations, we used different evaluation functions without tuning their hyper-parameters, but many others exist that can be adopted. Moreover, the whole process could be automated on the basis of collected data in order to provide a SEP-net that better describes a social behaviour.

Moreover, our model goes even further and provides a demonstration of how the current methods of implementing morality in AI systems would need to be modified to capture the computational mechanisms we describe.

7 Conclusions and future work

We have taken a first step to model and understand the question of when humans think it is morally acceptable to break rules. We showed that existing structures in the preference reasoning literature are insufficient for modeling this data, and we defined a generalization of CP-nets, called SEP-nets, which allow the linkage of preferences with scenarios and context evaluation. We constructed and studied a suite of hypothetical scenarios relating to this question, and collected human judgments about these scenarios.

Through our empirical study, we showed that humans seem to employ a complex set of preferences when determining if it is morally acceptable to break a previously established rule. Subjects seem to take into account the particular elements of a given scenario, including location and reasons to break the rule. This provides some evidence that a System 2 process is operative in rule-breaking decisions: implying that moral judgments were influenced by calculations of the impact of rule-violations on the well-being of others (consequentialist reasoning) as well as notions of what everyone would agree to (contractualist reasoning). Other times, System 1 reasoning seemed to be operative; sometimes, outcomes and agreement had no relationship to moral permissibility and rather rules were considered inviolable. Together, this pattern of data begins to suggest that moral rigidity and moral flexibility may be driven by fast and slow thinking.

There have been various attempts by philosophers to unite the three major threads of moral philosophy (outcomes, rules, and agreement) into a single unified view [86, 106]. Parfit called his unified view a “Triple Theory”. To date, no psychological theory has attempted to explain how rules, outcomes, and agreement are all integrated in the moral mind. Our work is one step on the way to creating a Psychological Triple Theory.

We do not claim that our proposal is the sole solution to the problem of how AI systems should reason about exceptions to constraints. We simply aim to describe one potential mechanism that humans use to reason about exceptions to one kind of rule (socio-moral rules) that could help AI systems do the same. Of course, we hope that the mechanism we describe here generalizes beyond the specific scenarios we use and perhaps even past the specific rule categories we use (e.g. to organizational norms, religious norms, social conventions, and so on). However, more extensive experimental and computational work will

be needed to determine how general this process is. Indeed, while the SEP-net formalism could take into account strict constraints in the form of scenario variables that engage in certain contexts, i.e., are not learned from data but rather set by experts or situational constraints. However, for sensitive applications such as health care, these would require careful understanding and modeling. Finally, we do not explicitly encode high-level moral values such as altruism or fairness in our model. While we could add a layer to the SEP-net, defining and quantifying these variables is beyond the scope of this paper and an interesting direction for future work. Among several possible extensions, we consider the following the most interesting.

7.1 Comparing preferences, measuring value deviation, and uncertainty

We defined generalized CP-nets (called SEP-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. Another fruitful direction for future work is to attempt to adapt our results to other preference reasoning formalisms, e.g., soft-constraints or weighted logical representations [52]. Reasoning about preferences and decision-making requires also being able to compare preferences. For instance, in a multi-agent system, it is important to understand whether an agent is deviating from a societal priority or norm. To do that, recent studies proposed metric spaces over preferences [69] which can be used to implement value alignment procedures. Such metric spaces may be extended to be applicable to SEP-nets. Such metrics would be useful to allow a system designer to intervene in cases where individuals (or artificial agents) behave differently than expected or are operating outside the norms or rules of a society. Finally, in our work so far we use deterministic formalisms for SEP-nets, however, moral situations are often complicated by uncertainty in the scenario or decisions made. A key future direction will be generalizing our work, e.g., through the use of probabilistic CP-nets (PCP-nets) [23] or other graphical utility models such as generalized random utility models [107], to explicitly model this uncertainty.

7.2 Limitations of our approach

While our work marks a significant step towards modeling and understanding the conditions under which humans find it morally acceptable to break the rules, it is important to acknowledge limitations in both our experiment and our model. In the scenario proposed in this work, the SEP-net focuses on a single target variable. However, real-world applications may require the consideration of multiple variables in the final decision-making stage. In such cases, the model would need to be equipped with additional preference variables to capture the complexity of these scenarios. This expansion could significantly increase the model's complexity, particularly with adding more variables per layer. Managing this complexity and ensuring the model remains computationally feasible is a critical challenge that needs further exploration.

Moreover, while our SEP-net framework provides a structured method to link preferences with scenarios and context evaluations, it does not fully address the complexity and uncertainty inherent in moral reasoning. The SEP-net formalism we define is extensible in that we can add additional layers without significantly affecting the semantics of the current formalism. However, given that identifying and quantifying these values, especially in the context of moral decision making, is an active area of research in moral theory significant work is needed to model these.

Finally, in some settings we may wish to incorporate expert knowledge or constraints into the model. While the SEP-net framework as defined should be able to handle these issues, the semantics are not straightforward. Likewise, the current SEP-net framework could be extended to incorporate probabilistic elements, handle multiple target variables, and manage the increased complexity from additional variables to enhance the robustness and applicability of our model.

7.3 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 in creating plans, consisting of sequences of actions, that would respect or follow certain preferences [108]. 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 [109]. 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 [110]. 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 [111]. 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.

Appendix A Complete statistical tables

See Tables 8, 9 and 10.

Table 8 p-values for the Wilcoxon signed-rank test against the null-hypothesis: changing the order of evaluation or preference does not influence the preference value

Scenario	PREFthenEVAL	PREFthenEVAL	p-value	Rejected
DFO_SP	0.1212 (0.3264)	0.1692 (0.3750)	0.5059	False
DFO_WA	0.4697 (0.4991)	0.4154 (0.4928)	0.5860	False
DFO_SO	0.6970 (0.4596)	0.6923 (0.4615)	0.8535	False
DFO_CA	0.7576 (0.4285)	0.6154 (0.4865)	0.0743	False
DFO_CO	0.7424 (0.4373)	0.6308 (0.4826)	0.1979	False
DFO_SU	0.0909 (0.2875)	0.1385 (0.3454)	0.4281	False
DFO_BR	0.0758 (0.2646)	0.1692 (0.3750)	0.1414	False
DFO_HS	0.1061 (0.3079)	0.1385 (0.3454)	0.6379	False
DFO_TP	0.2121 (0.4088)	0.1692 (0.3750)	0.4579	False
DFO_MA	0.2879 (0.4528)	0.3231 (0.4677)	0.5860	False
DFO_FA	0.2424 (0.4285)	0.3231 (0.4677)	0.3596	False
DFO_SW	0.8182 (0.3857)	0.7538 (0.4308)	0.4236	False
BFO_HA	0.5507 (0.4974)	0.5217 (0.4995)	0.7317	False
BFO_CL	0.1594 (0.3661)	0.2899 (0.4537)	0.0971	False
BFO_TU	0.1014 (0.3019)	0.1884 (0.3910)	0.1414	False
BFO_JA	0.1594 (0.3661)	0.2319 (0.4220)	0.3248	False
BFO_FR	0.7101 (0.4537)	0.7536 (0.4309)	0.5730	False
BFO_EL	0.1014 (0.3019)	0.2029 (0.4022)	0.0948	False
BFO_BR	0.7391 (0.4391)	0.7681 (0.4220)	0.7238	False
AFO_MN	0.4030 (0.4905)	0.4219 (0.4939)	0.6096	False
AFO_BA	0.4328 (0.4955)	0.4531 (0.4978)	0.7526	False
AFO_JA	0.4478 (0.4973)	0.5625 (0.4961)	0.1088	False
AFO_CA	0.5224 (0.4995)	0.5781 (0.4939)	0.3471	False
AFO_BR	0.4478 (0.4973)	0.4062 (0.4911)	0.4285	False
AFO_HR	0.8209 (0.3834)	0.7969 (0.4023)	0.5607	False

The null-hypothesis is rejected if $p < 0.05$

Table 9 p-values for the Wilcoxon signed-rank test against the null-hypothesis NH4: evaluation variable does not affect the preference

Scenario	EV	Group 0–1	Group 0–2	Group 0–3	Group 1–2	Group 1–3	Group 2–3
DFO_SP	everyone	0.1489	0.4237	0.4840	1.0000	0.7656	0.7656
DFO_WA	everyone	0.3506	0.3458	**0.0004**	1.0000	**0.0050**	0.3458
DFO_SO	everyone	1.0000	0.3458	**0.0001**	0.7728	**0.0001**	**0.0369**
DFO_CA	everyone	0.1817	0.0726	**0.0000**	0.1817	**0.0001**	0.1096
DFO_CO	everyone	0.3506	0.0719	**0.0000**	0.0719	**0.0000**	nan
DFO_SU	everyone	0.2986	0.7728	**0.0369**	1.0000	0.4237	1.0000
DFO_BR	everyone	0.2330	1.0000	1.0000	1.0000	0.7897	0.7656
DFO_HS	everyone	0.1489	1.0000	0.7728	0.1489	0.7656	0.0719
DFO_TP	everyone	0.1169	**0.0411**	**0.0077**	0.7728	0.2986	0.7728
DFO_MA	everyone	**0.0026**	0.7768	**0.0219**	0.1489	0.5653	0.4840
DFO_FA	everyone	**0.0048**	0.3458	**0.0019**	0.4237	0.7768	0.0719
DFO_SW	everyone	1.0000	**0.0060**	**0.0001**	**0.0060**	**0.0006**	0.7656
BFO_HA	everyone	**0.0134**	**0.0004**	**0.0000**	0.0572	**0.0030**	0.7897
BFO_CL	everyone	0.4835	**0.0355**	**0.0003**	0.3506	**0.0107**	0.2330
BFO_TU	everyone	0.0658	0.7728	0.8016	1.0000	0.1817	1.0000
BFO_JA	everyone	0.7728	0.7656	0.0719	0.5653	0.7768	0.1096
BFO_FR	everyone	**0.0197**	**0.0369**	**0.0000**	0.0719	**0.0000**	**0.0197**
BFO_EL	everyone	**0.0369**	0.3458	0.5653	1.0000	0.1450	1.0000
BFO_BR	everyone	0.2330	0.4840	**0.0000**	0.4840	**0.0002**	0.0726
AFO_MN	everyone	0.8293	1.0000	0.0628	0.4840	**0.0140**	0.4237
AFO_BA	everyone	0.4579	0.1779	**0.0005**	0.5653	**0.0019**	**0.0411**
AFO_JA	everyone	0.1779	0.1169	**0.0030**	0.5653	0.0572	0.4281
AFO_CA	everyone	**0.0135**	0.4840	**0.0002**	0.7728	**0.0235**	0.4840
AFO_BR	everyone	0.0572	**0.0488**	**0.0411**	0.8016	0.5941	0.7768
AFO_HR	everyone	0.1489	1.0000	**0.0000**	1.0000	**0.0000**	0.0719
DFO_SP	first_person	0.7656	1.0000	1.0000	0.0915	0.2273	0.7656
DFO_WA	first_person	0.5297	0.4281	**0.0026**	0.1647	**0.0006**	0.0572
DFO_SO	first_person	0.3506	0.4840	**0.0001**	1.0000	**0.0004**	**0.0234**
DFO_CA	first_person	0.1489	**0.0006**	**0.0001**	**0.0235**	**0.0026**	0.7768
DFO_CO	first_person	1.0000	0.1450	**0.0000**	0.1817	**0.0000**	0.0658
DFO_SU	first_person	0.3506	**0.0411**	**0.0411**	0.4237	0.4840	0.7728
DFO_BR	first_person	0.7728	0.7656	0.1294	0.3458	0.1096	0.2986
DFO_HS	first_person	0.5297	0.1294	0.4840	0.4237	1.0000	0.2330
DFO_TP	first_person	**0.0050**	**0.0004**	**0.0006**	0.7728	0.5297	1.0000
DFO_MA	first_person	**0.0412**	**0.0083**	**0.0034**	0.2669	0.6179	0.7897
DFO_FA	first_person	0.4281	**0.0006**	**0.0234**	0.1096	0.1817	0.7656
DFO_SW	first_person	nan	**0.0107**	**0.0000**	**0.0107**	**0.0000**	0.1450
BFO_HA	first_person	**0.0234**	**0.0000**	**0.0000**	**0.0002**	**0.0019**	1.0000
BFO_CL	first_person	0.5941	**0.0077**	**0.0002**	0.1294	0.2986	1.0000
BFO_TU	first_person	**0.0394**	0.1294	0.8016	0.7728	0.1817	0.1294
BFO_JA	first_person	0.7728	0.3458	0.7728	0.8016	0.3929	0.5941
BFO_FR	first_person	0.7728	**0.0394**	**0.0000**	**0.0019**	**0.0015**	**0.0051**
BFO_EL	first_person	**0.0197**	**0.0411**	0.5297	0.7728	0.2986	**0.0411**
BFO_BR	first_person	0.2986	0.1489	**0.0000**	0.7728	**0.0002**	**0.0411**

Table 9 (continued)

Scenario	EV	Group 0–1	Group 0–2	Group 0–3	Group 1–2	Group 1–3	Group 2–3
AFO_MN	first_person	0.2081	0.7768	**0.0311**	1.0000	**0.0009**	**0.0234**
AFO_BA	first_person	0.8121	**0.0235**	**0.0001**	**0.0084**	**0.0002**	**0.0019**
AFO_JA	first_person	0.3506	0.0915	**0.0219**	0.0759	**0.0003**	**0.0197**
AFO_CA	first_person	0.1779	0.0658	**0.0005**	1.0000	**0.0124**	**0.0411**
AFO_BR	first_person	0.6179	**0.0134**	0.6179	**0.0048**	0.1169	**0.0488**
AFO_HR	first_person	nan	0.0719	**0.0000**	0.2330	**0.0001**	**0.0018**
DFO_SP	middle_person	0.4237	0.7656	0.4237	**0.0411**	0.1096	0.4237
DFO_WA	middle_person	0.7768	**0.0235**	**0.0011**	0.2357	**0.0011**	0.1779
DFO_SO	middle_person	1.0000	**0.0234**	**0.0000**	**0.0411**	**0.0000**	0.1450
DFO_CA	middle_person	0.2330	**0.0060**	**0.0000**	**0.0394**	**0.0002**	0.7656
DFO_CO	middle_person	1.0000	0.3458	**0.0000**	0.7728	**0.0000**	0.0719
DFO_SU	middle_person	0.2330	0.2330	0.2330	0.7728	0.0726	0.7728
DFO_BR	middle_person	0.5297	0.7728	0.7897	1.0000	0.5297	0.2986
DFO_HS	middle_person	1.0000	**0.0369**	**0.0369**	**0.0369**	0.2330	0.2330
DFO_TP	middle_person	0.0915	**0.0219**	**0.0140**	0.3506	0.0726	1.0000
DFO_MA	middle_person	**0.0026**	**0.0030**	**0.0268**	0.5297	0.8213	0.5653
DFO_FA	middle_person	**0.0083**	**0.0107**	**0.0135**	0.1294	0.7768	0.7728
DFO_SW	middle_person	0.7728	0.1489	**0.0001**	0.1489	**0.0000**	**0.0369**
BFO_HA	middle_person	0.5297	**0.0002**	**0.0004**	**0.0001**	**0.0000**	0.8016
BFO_CL	middle_person	0.5941	**0.0004**	**0.0009**	0.1096	0.0726	1.0000
BFO_TU	middle_person	**0.0411**	0.1817	0.1779	0.7656	0.4840	1.0000
BFO_JA	middle_person	0.2330	0.7728	1.0000	0.2081	0.2669	0.4840
BFO_FR	middle_person	0.4237	**0.0134**	**0.0000**	**0.0394**	**0.0000**	0.0915
BFO_EL	middle_person	0.4237	**0.0369**	1.0000	0.7728	0.1450	0.1294
BFO_BR	middle_person	0.7656	**0.0369**	**0.0000**	0.2330	**0.0001**	0.0658
AFO_MN	middle_person	0.4281	0.8121	**0.0009**	1.0000	**0.0140**	**0.0235**
AFO_BA	middle_person	0.2357	**0.0107**	**0.0005**	0.2357	**0.0051**	**0.0084**
AFO_JA	middle_person	0.0572	**0.0219**	**0.0001**	0.8016	**0.0011**	**0.0077**
AFO_CA	middle_person	**0.0488**	**0.0355**	**0.0002**	0.8213	**0.0140**	**0.0077**
AFO_BR	middle_person	0.4579	**0.0235**	**0.0235**	**0.0005**	0.0916	0.2273
AFO_HR	middle_person	1.0000	0.1489	**0.0000**	0.1489	**0.0000**	**0.0034**
DFO_SP	last_person	0.2330	0.1489	0.1489	0.7656	1.0000	1.0000
DFO_WA	last_person	0.2669	0.0719	**0.0050**	0.7728	**0.0077**	1.0000
DFO_SO	last_person	0.7897	0.1489	**0.0000**	0.3458	**0.0001**	1.0000
DFO_CA	last_person	0.1096	**0.0369**	**0.0000**	0.4237	**0.0003**	0.3458
DFO_CO	last_person	1.0000	0.1489	**0.0000**	0.1489	**0.0000**	1.0000
DFO_SU	last_person	**0.0134**	1.0000	**0.0134**	1.0000	0.2330	nan
DFO_BR	last_person	0.7897	0.7728	1.0000	0.7728	0.7897	1.0000
DFO_HS	last_person	**0.0107**	**0.0411**	0.1096	1.0000	0.4840	0.7728
DFO_TP	last_person	0.1450	**0.0060**	**0.0394**	0.1096	0.3506	0.7656
DFO_MA	last_person	**0.0124**	0.2330	**0.0051**	0.4237	0.5297	0.2330
DFO_FA	last_person	**0.0011**	0.0726	**0.0030**	1.0000	0.7768	1.0000
DFO_SW	last_person	0.3458	**0.0369**	**0.0000**	**0.0369**	**0.0002**	0.7728
BFO_HA	last_person	**0.0411**	**0.0077**	**0.0001**	0.2669	**0.0124**	0.7656

Table 9 (continued)

Scenario	EV	Group 0–1	Group 0–2	Group 0–3	Group 1–2	Group 1–3	Group 2–3
BFO_CL	last_person	0.0572	**0.0045**	**0.0083**	0.2986	0.3506	0.7656
BFO_TU	last_person	0.8016	0.4840	0.5653	1.0000	0.8016	1.0000
BFO_JA	last_person	0.4237	1.0000	0.4237	0.1096	0.3506	0.4237
BFO_FR	last_person	**0.0197**	**0.0107**	**0.0000**	**0.0197**	**0.0004**	0.1294
BFO_EL	last_person	0.1817	0.5297	0.2986	0.1817	0.2986	0.3506
BFO_BR	last_person	1.0000	0.3458	**0.0000**	1.0000	**0.0001**	0.2330
AFO_MN	last_person	0.5941	0.2330	**0.0084**	0.4840	**0.0015**	0.1294
AFO_BA	last_person	0.1414	0.3929	**0.0005**	0.7656	**0.0124**	**0.0369**
AFO_JA	last_person	0.2081	0.1294	**0.0019**	0.4237	0.0519	0.2330
AFO_CA	last_person	**0.0289**	0.1489	**0.0219**	1.0000	0.0948	0.4237
AFO_BR	last_person	0.8121	0.3506	0.6179	1.0000	1.0000	0.7897
AFO_HR	last_person	0.3458	0.0719	**0.0000**	0.0719	**0.0000**	**0.0077**
DFO_SP	cutter	0.7656	0.7897	0.7897	0.5297	0.7897	1.0000
DFO_WA	cutter	0.0759	0.3014	0.8121	0.6179	0.2273	0.3318
DFO_SO	cutter	1.0000	0.0915	0.7768	**0.0355**	0.8016	0.1294
DFO_CA	cutter	0.1096	0.8016	**0.0134**	0.8121	0.2669	0.1779
DFO_CO	cutter	0.1450	0.3506	0.7768	0.4835	**0.0197**	0.1450
DFO_SU	cutter	0.7728	0.4237	0.1489	0.4840	0.2330	0.7728
DFO_BR	cutter	0.7897	0.2986	0.4237	0.7768	0.2330	0.0726
DFO_HS	cutter	0.7728	0.2330	1.0000	0.7656	1.0000	1.0000
DFO_TP	cutter	0.5653	0.1294	0.5941	1.0000	1.0000	1.0000
DFO_MA	cutter	0.2669	0.4281	0.5653	0.7897	0.6179	0.6179
DFO_FA	cutter	0.8213	0.2273	0.7897	1.0000	0.2273	0.3014
DFO_SW	cutter	0.7656	0.7897	1.0000	0.8121	0.8121	0.8016
BFO_HA	cutter	0.0759	0.1779	0.3826	0.6179	0.5941	0.5941
BFO_CL	cutter	0.4281	0.2330	0.5941	0.2986	0.5297	0.1817
BFO_TU	cutter	0.1294	0.5297	0.4840	0.4840	1.0000	0.2330
BFO_JA	cutter	0.8213	0.1096	0.1817	0.0915	0.2273	0.0719
BFO_FR	cutter	0.6179	0.3014	0.3929	0.2357	0.5941	0.0948
BFO_EL	cutter	0.4281	0.4840	0.2669	0.4237	1.0000	1.0000
BFO_BR	cutter	**0.0034**	0.2330	0.4237	**0.0394**	**0.0011**	**0.0134**
AFO_MN	cutter	0.2669	**0.0084**	**0.0394**	0.1169	0.8121	0.1096
AFO_BA	cutter	0.5653	0.5297	1.0000	0.8016	0.4835	0.3318
AFO_JA	cutter	0.3248	0.0658	0.3014	1.0000	0.1169	0.0628
AFO_CA	cutter	**0.0235**	**0.0311**	0.4281	0.4835	1.0000	0.1779
AFO_BR	cutter	**0.0412**	0.3318	**0.0134**	0.5653	0.3506	0.4281
AFO_HR	cutter	**0.0034**	**0.0011**	0.3458	0.0572	**0.0394**	**0.0045**
DFO_SP	univers.	0.1489	0.4237	0.1489	0.5941	0.2273	0.2986
DFO_WA	univers.	0.0726	**0.0015**	**0.0018**	**0.0346**	**0.0045**	0.3929
DFO_SO	univers.	0.2986	**0.0369**	**0.0000**	0.7768	**0.0000**	**0.0077**
DFO_CA	univers.	**0.0411**	**0.0060**	**0.0000**	**0.0060**	**0.0018**	0.2986
DFO_CO	univers.	0.4237	**0.0019**	**0.0000**	**0.0107**	**0.0000**	0.5297
DFO_SU	univers.	0.7728	0.3458	0.7728	0.1817	0.1817	1.0000
DFO_BR	univers.	0.7768	0.4840	0.7897	0.7656	0.7768	0.7768

Table 9 (continued)

Scenario	EV	Group 0–1	Group 0–2	Group 0–3	Group 1–2	Group 1–3	Group 2–3
DFO_HS	univers.	0.2330	nan	1.0000	0.2330	0.3506	0.7728
DFO_TP	univers.	**0.0411**	0.0726	**0.0234**	0.1817	0.1450	1.0000
DFO_MA	univers.	0.0572	**0.0140**	**0.0015**	0.1169	0.0658	0.7656
DFO_FA	univers.	0.1134	**0.0034**	**0.0124**	**0.0411**	0.2669	0.7656
DFO_SW	univers.	0.7728	**0.0060**	**0.0002**	**0.0060**	**0.0011**	0.7768
BFO_HA	univers.	0.8213	**0.0078**	**0.0004**	**0.0030**	**0.0003**	0.7656
BFO_CL	univers.	0.4281	**0.0018**	**0.0005**	0.1096	**0.0411**	1.0000
BFO_TU	univers.	1.0000	0.0726	0.5297	**0.0411**	0.7768	0.1294
BFO_JA	univers.	0.3506	0.2330	0.7656	0.1779	1.0000	1.0000
BFO_FR	univers.	0.1489	**0.0006**	**0.0000**	**0.0140**	**0.0001**	**0.0311**
BFO_EL	univers.	0.7728	0.7768	0.4237	1.0000	0.8016	0.7768
BFO_BR	univers.	0.3458	**0.0197**	**0.0000**	0.1294	**0.0003**	0.0658
AFO_MN	univers.	1.0000	0.2986	**0.0355**	1.0000	**0.0140**	0.1817
AFO_BA	univers.	0.4281	**0.0084**	**0.0050**	**0.0355**	**0.0048**	0.7897
AFO_JA	univers.	0.5941	0.0628	**0.0011**	0.1414	**0.0015**	0.1450
AFO_CA	univers.	0.0948	**0.0011**	**0.0019**	**0.0197**	0.1779	0.3929
AFO_BR	univers.	**0.0394**	**0.0001**	**0.0140**	**0.0011**	0.0948	0.1450
AFO_HR	univers.	0.3458	0.1489	**0.0000**	0.4237	**0.0005**	**0.0135**
DFO_SP	likelihood	0.0726	**0.0197**	0.1817	0.7656	0.4840	0.4840
DFO_WA	likelihood	1.0000	0.7897	0.0726	0.8213	0.0572	0.1662
DFO_SO	likelihood	0.7728	0.4840	**0.0034**	0.5941	0.2081	0.1096
DFO_CA	likelihood	0.7897	**0.0135**	**0.0015**	**0.0355**	**0.0050**	0.1450
DFO_CO	likelihood	0.7728	**0.0009**	**0.0001**	**0.0011**	**0.0001**	**0.0107**
DFO_SU	likelihood	0.3506	0.7728	1.0000	0.4237	0.7656	1.0000
DFO_BR	likelihood	1.0000	0.7728	1.0000	1.0000	1.0000	0.7656
DFO_HS	likelihood	0.4840	1.0000	1.0000	0.5297	0.7768	0.7768
DFO_TP	likelihood	0.8213	**0.0197**	0.0726	0.0915	0.2669	0.7768
DFO_MA	likelihood	**0.0015**	**0.0006**	**0.0004**	0.7768	0.5653	0.8016
DFO_FA	likelihood	**0.0107**	**0.0045**	**0.0009**	0.1450	0.2273	0.8016
DFO_SW	likelihood	0.0719	**0.0034**	**0.0001**	0.1096	**0.0045**	0.2273
BFO_HA	likelihood	0.2081	0.0771	**0.0124**	0.5059	0.2357	0.8213
BFO_CL	likelihood	0.3929	**0.0083**	**0.0083**	0.3506	0.1294	0.7768
BFO_TU	likelihood	0.5297	0.3929	1.0000	0.7768	0.4840	0.4840
BFO_JA	likelihood	0.2669	0.3929	0.1450	0.1817	1.0000	0.2273
BFO_FR	likelihood	0.7656	0.1294	**0.0077**	0.2669	0.0630	0.2610
BFO_EL	likelihood	0.3506	0.4840	0.1096	0.2986	0.5297	0.5297
BFO_BR	likelihood	0.7656	**0.0005**	**0.0077**	**0.0009**	**0.0077**	1.0000
AFO_MN	likelihood	0.1817	**0.0045**	0.2081	0.0948	0.6379	0.2669
AFO_BA	likelihood	0.2273	**0.0002**	**0.0030**	**0.0048**	**0.0488**	0.8121
AFO_JA	likelihood	0.2669	**0.0031**	**0.0006**	**0.0135**	**0.0015**	0.2081
AFO_CA	likelihood	0.3014	**0.0311**	**0.0007**	0.2357	**0.0051**	0.2669
AFO_BR	likelihood	**0.0140**	**0.0219**	**0.0234**	0.6379	0.8016	1.0000
AFO_HR	likelihood	0.1489	**0.0034**	**0.0004**	**0.0411**	**0.0019**	0.0572

The null-hypothesis is rejected if $p < 0.05$, these cases are reported in bold. Note that the NaN values are for elements where there is not enough support to perform the test but do not affect our overall results

Table 10 p-values for the Wilcoxon signed-rank test against the null-hypothesis NH2: reason does not affect evaluation variables

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
DFO_SP	DFO_WA	0.2576	0.2927	0.0810	0.1037	0.5957	**0.0393**	0.9389
DFO_SP	DFO_SO	**0.0098**	**0.0238**	**0.0050**	**0.0029**	0.1911	**0.0007**	0.4266
DFO_SP	DFO_CA	**0.0001**	**0.0005**	**0.0000**	**0.0000**	0.1741	**0.0000**	**0.0000**
DFO_SP	DFO_CO	**0.0026**	**0.0117**	**0.0023**	**0.0020**	**0.0083**	**0.0033**	**0.0000**
DFO_SP	DFO_SU	0.1426	0.1960	0.1692	**0.0200**	**0.0000**	0.1937	**0.0000**
DFO_SP	DFO_BR	0.1524	0.0510	0.0742	0.0533	0.6433	0.6679	**0.0000**
DFO_SP	DFO_HS	0.6004	0.9483	0.9699	0.6789	**0.0006**	0.6405	**0.0120**
DFO_SP	DFO_TP	0.1848	0.3437	0.2102	0.0895	0.8971	0.6394	**0.0072**
DFO_SP	DFO_MA	**0.0372**	0.2239	0.0545	**0.0270**	0.3939	**0.0150**	0.8042
DFO_SP	DFO_FA	**0.0041**	**0.0051**	**0.0002**	**0.0004**	0.7502	**0.0003**	0.6971
DFO_SP	DFO_SW	**0.0001**	**0.0001**	**0.0006**	**0.0001**	0.0800	**0.0000**	**0.0000**
DFO_SP	BFO_HA	0.1429	**0.0113**	0.0872	0.1495	0.8587	**0.0080**	**0.0003**
DFO_SP	BFO_CL	0.1021	0.0826	**0.0134**	**0.0279**	**0.0016**	0.4187	**0.0015**
DFO_SP	BFO_TU	0.4174	**0.0014**	**0.0265**	0.7002	**0.0020**	0.4160	**0.0001**
DFO_SP	BFO_JA	0.9744	0.4098	0.4025	0.6791	0.8361	0.0838	**0.0019**
DFO_SP	BFO_FR	**0.0002**	**0.0299**	**0.0002**	**0.0372**	0.1481	**0.0000**	**0.0160**
DFO_SP	BFO_EL	0.7901	0.1970	0.0631	0.5477	0.2458	0.4133	**0.0010**
DFO_SP	BFO_BR	**0.0003**	**0.0001**	**0.0003**	**0.0103**	**0.0412**	**0.0000**	**0.0000**
DFO_SP	AFO_MN	**0.0005**	**0.0162**	**0.0055**	**0.0239**	**0.0143**	**0.0024**	0.0526
DFO_SP	AFO_BA	0.4057	0.6631	0.1543	0.4234	0.0817	0.1124	**0.0007**
DFO_SP	AFO_JA	**0.0225**	0.1121	**0.0223**	**0.0276**	0.7590	**0.0074**	**0.0004**
DFO_SP	AFO_CA	**0.0109**	0.2094	**0.0005**	**0.0007**	0.3199	**0.0009**	**0.0024**
DFO_SP	AFO_BR	0.1522	0.3195	0.1226	0.2088	0.2656	0.0529	**0.0002**
DFO_SP	AFO_HR	**0.0006**	**0.0201**	**0.0004**	**0.0005**	0.1467	**0.0000**	**0.0000**
DFO_WA	DFO_SO	0.0809	0.2738	0.2395	0.2102	0.4666	0.3133	0.3244
DFO_WA	DFO_CA	**0.0012**	**0.0057**	**0.0004**	**0.0052**	0.3893	**0.0387**	**0.0000**
DFO_WA	DFO_CO	**0.0193**	0.0547	**0.0377**	0.0678	0.0635	0.3542	**0.0000**
DFO_WA	DFO_SU	0.4977	0.9118	0.8108	0.6923	**0.0000**	0.1857	**0.0000**
DFO_WA	DFO_BR	0.8102	0.4233	0.3115	0.4445	0.8805	0.2035	**0.0000**
DFO_WA	DFO_HS	0.2931	0.3330	0.3982	0.1198	**0.0000**	**0.0024**	**0.0016**
DFO_WA	DFO_TP	0.6326	0.9119	0.7483	0.7279	0.5474	0.1665	**0.0008**
DFO_WA	DFO_MA	0.2616	0.8089	0.3557	0.3847	0.7136	0.5187	0.9242
DFO_WA	DFO_FA	**0.0234**	**0.0219**	**0.0109**	**0.0399**	0.6142	0.0931	0.5696
DFO_WA	DFO_SW	**0.0010**	**0.0094**	**0.0030**	**0.0117**	0.3936	**0.0061**	**0.0000**
DFO_WA	BFO_HA	0.4243	0.2953	0.6165	0.9080	0.7770	0.4610	**0.0001**
DFO_WA	BFO_CL	0.2198	0.3432	0.0710	0.3384	**0.0006**	0.3357	**0.0020**
DFO_WA	BFO_TU	0.8157	**0.0393**	0.0792	0.2403	**0.0009**	0.5588	**0.0000**
DFO_WA	BFO_JA	0.4414	0.2188	0.5775	**0.0281**	0.1790	0.5868	**0.0011**
DFO_WA	BFO_FR	**0.0015**	0.0790	**0.0040**	0.2848	0.3623	**0.0033**	**0.0103**
DFO_WA	BFO_EL	0.5926	**0.0305**	**0.0017**	0.1431	**0.0320**	0.2020	**0.0000**
DFO_WA	BFO_BR	**0.0010**	**0.0004**	**0.0028**	0.3873	**0.0366**	**0.0029**	**0.0000**
DFO_WA	AFO_MN	**0.0247**	0.2322	0.1289	0.3033	**0.0340**	0.1198	**0.0495**
DFO_WA	AFO_BA	0.9641	0.7808	0.5697	0.5959	0.2882	0.6050	**0.0003**

Table 10 (continued)

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
DFO_WA	AFO_JA	0.0627	0.9041	0.3082	0.5442	0.2743	0.4018	**0.0000**
DFO_WA	AFO_CA	0.0603	0.7259	0.1034	0.0662	0.8009	0.3581	**0.0011**
DFO_WA	AFO_BR	0.5243	0.5340	0.6524	0.6965	0.0769	1.0000	**0.0000**
DFO_WA	AFO_HR	**0.0061**	0.1679	**0.0275**	0.0744	0.0703	**0.0278**	**0.0000**
DFO_SO	DFO_CA	0.0535	**0.0364**	**0.0324**	**0.0303**	0.9677	0.1498	**0.0001**
DFO_SO	DFO_CO	0.4275	0.4653	0.4120	0.4079	0.1008	0.9270	**0.0000**
DFO_SO	DFO_SU	0.3132	0.2521	0.6091	0.2707	**0.0000**	**0.0336**	**0.0000**
DFO_SO	DFO_BR	0.3188	0.9848	0.6210	0.4240	0.6132	**0.0149**	**0.0000**
DFO_SO	DFO_HS	**0.0053**	**0.0445**	**0.0096**	**0.0157**	**0.0000**	**0.0002**	**0.0449**
DFO_SO	DFO_TP	0.0763	0.4105	0.2800	0.3020	0.2503	**0.0098**	**0.0453**
DFO_SO	DFO_MA	0.7235	0.6324	0.6505	0.7717	0.5832	0.7200	0.2890
DFO_SO	DFO_FA	0.2642	0.4386	0.0935	0.2806	0.5107	0.4452	0.4994
DFO_SO	DFO_SW	**0.0150**	**0.0214**	**0.0306**	0.1591	0.6057	**0.0338**	**0.0000**
DFO_SO	BFO_HA	0.4831	0.9282	0.2899	0.1298	0.4661	0.9400	**0.0049**
DFO_SO	BFO_CL	0.8919	0.8587	0.6538	0.9365	**0.0002**	0.1602	**0.0233**
DFO_SO	BFO_TU	0.1491	0.1689	0.8089	0.0874	**0.0212**	0.2677	**0.0004**
DFO_SO	BFO_JA	**0.0147**	0.0516	0.0739	**0.0031**	0.0988	0.1454	**0.0109**
DFO_SO	BFO_FR	0.0765	0.9372	0.0817	0.8678	0.6543	0.0787	**0.0471**
DFO_SO	BFO_EL	**0.0363**	**0.0106**	**0.0000**	**0.0295**	**0.0156**	**0.0327**	**0.0095**
DFO_SO	BFO_BR	**0.0189**	**0.0370**	0.3002	0.8900	0.1013	**0.0219**	**0.0003**
DFO_SO	AFO_MN	0.5799	0.9279	0.7795	0.7396	0.0504	0.8105	0.2208
DFO_SO	AFO_BA	0.1235	0.2341	0.2139	0.1560	0.5754	0.1070	**0.0094**
DFO_SO	AFO_JA	0.9093	0.3036	0.7394	0.7887	0.1056	0.9334	**0.0123**
DFO_SO	AFO_CA	0.7481	0.4742	0.7371	0.5164	0.9396	0.8671	**0.0259**
DFO_SO	AFO_BR	0.1693	0.2174	0.3793	0.1020	**0.0317**	0.3568	**0.0045**
DFO_SO	AFO_HR	0.1154	0.3854	0.4894	0.5011	**0.0312**	0.1873	**0.0000**
DFO_CA	DFO_CO	0.2840	0.1993	0.2225	0.1837	0.2728	0.1201	**0.0004**
DFO_CA	DFO_SU	**0.0087**	**0.0057**	**0.0011**	**0.0132**	**0.0001**	**0.0008**	**0.0055**
DFO_CA	DFO_BR	**0.0075**	0.1540	**0.0114**	**0.0221**	0.3660	**0.0007**	0.5218
DFO_CA	DFO_HS	**0.0000**	**0.0012**	**0.0001**	**0.0001**	**0.0000**	**0.0000**	**0.0112**
DFO_CA	DFO_TP	**0.0020**	**0.0116**	**0.0020**	**0.0051**	0.3191	**0.0003**	**0.0292**
DFO_CA	DFO_MA	**0.0304**	**0.0109**	**0.0044**	**0.0143**	0.6385	**0.0315**	**0.0000**
DFO_CA	DFO_FA	0.4449	0.6688	0.2591	0.2504	0.5822	0.3860	**0.0000**
DFO_CA	DFO_SW	0.9393	0.6834	0.4835	0.7528	0.5466	0.8922	0.4865
DFO_CA	BFO_HA	**0.0349**	**0.0365**	**0.0018**	**0.0005**	0.4741	0.1456	0.1777
DFO_CA	BFO_CL	0.0599	0.1245	**0.0347**	**0.0371**	**0.0002**	**0.0087**	**0.0445**
DFO_CA	BFO_TU	**0.0049**	0.3064	**0.0219**	**0.0002**	**0.0480**	**0.0163**	0.4576
DFO_CA	BFO_JA	**0.0001**	**0.0012**	**0.0002**	**0.0000**	0.2330	**0.0117**	**0.0306**
DFO_CA	BFO_FR	0.6946	0.1611	0.2320	0.1063	0.7065	0.8919	**0.0019**
DFO_CA	BFO_EL	**0.0003**	**0.0002**	**0.0000**	**0.0001**	**0.0186**	**0.0003**	0.1268
DFO_CA	BFO_BR	0.7582	0.8718	0.4569	**0.0154**	0.1262	0.5079	0.9991
DFO_CA	AFO_MN	0.1564	0.0912	**0.0299**	**0.0142**	0.1866	0.2167	**0.0011**
DFO_CA	AFO_BA	**0.0042**	**0.0093**	**0.0029**	**0.0006**	0.5638	**0.0048**	**0.0322**
DFO_CA	AFO_JA	0.1055	**0.0162**	**0.0079**	**0.0152**	0.1015	0.0906	0.0965

Table 10 (continued)

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
DFO_CA	AFO_CA	0.1097	**0.0169**	0.0658	0.1182	0.9431	0.2823	0.0523
DFO_CA	AFO_BR	**0.0023**	**0.0034**	**0.0031**	**0.0004**	**0.0359**	**0.0141**	0.0571
DFO_CA	AFO_HR	0.7626	0.1247	0.0797	0.1343	**0.0067**	0.7590	0.2729
DFO_CO	DFO_SU	0.0949	0.0956	**0.0365**	0.0808	**0.0072**	**0.0424**	0.2502
DFO_CO	DFO_BR	**0.0365**	0.7662	0.1419	0.2085	**0.0409**	**0.0145**	**0.0139**
DFO_CO	DFO_HS	**0.0013**	**0.0144**	**0.0057**	**0.0052**	**0.0000**	**0.0030**	**0.0000**
DFO_CO	DFO_TP	0.0662	0.1088	0.1125	0.0683	**0.0317**	**0.0083**	**0.0000**
DFO_CO	DFO_MA	0.1873	0.2092	0.0853	0.2264	0.0942	0.4718	**0.0000**
DFO_CO	DFO_FA	0.8298	0.8334	0.8426	0.6798	**0.0238**	0.6041	**0.0000**
DFO_CO	DFO_SW	0.3395	0.2520	0.4734	0.6084	0.4423	0.0516	**0.0041**
DFO_CO	BFO_HA	0.0916	0.2928	0.0608	0.0560	**0.0293**	0.8990	**0.0000**
DFO_CO	BFO_CL	0.2265	0.7764	0.8040	0.4438	**0.0000**	0.1502	**0.0000**
DFO_CO	BFO_TU	**0.0467**	0.8564	0.1806	**0.0117**	0.1374	0.3126	**0.0000**
DFO_CO	BFO_JA	**0.0021**	**0.0147**	**0.0138**	**0.0006**	**0.0092**	0.1365	**0.0000**
DFO_CO	BFO_FR	0.4887	0.2877	0.8508	0.3308	0.1592	0.0728	**0.0000**
DFO_CO	BFO_EL	**0.0186**	**0.0003**	**0.0000**	**0.0013**	**0.0011**	**0.0371**	**0.0000**
DFO_CO	BFO_BR	0.4100	0.1199	0.4167	0.6674	0.9771	0.0749	**0.0000**
DFO_CO	AFO_MN	0.8771	0.3757	0.4481	0.2101	0.8702	0.9670	**0.0000**
DFO_CO	AFO_BA	**0.0321**	**0.0363**	0.0703	**0.0413**	0.3614	0.2996	**0.0000**
DFO_CO	AFO_JA	0.4268	0.1705	0.1705	0.1330	**0.0016**	0.8324	**0.0000**
DFO_CO	AFO_CA	0.5125	0.2991	0.6100	0.8659	0.1423	0.8928	**0.0000**
DFO_CO	AFO_BR	0.0667	**0.0213**	0.1008	**0.0098**	**0.0007**	0.3860	**0.0000**
DFO_CO	AFO_HR	0.7014	0.8485	0.5030	0.7670	**0.0005**	0.2446	**0.0172**
DFO_SU	DFO_BR	0.7235	0.3874	0.5418	0.8906	**0.0000**	0.8361	0.1686
DFO_SU	DFO_HS	0.0541	0.3312	0.1801	0.1098	**0.0000**	0.1147	**0.0000**
DFO_SU	DFO_TP	0.8582	0.8737	0.8659	0.7367	**0.0000**	0.6684	**0.0000**
DFO_SU	DFO_MA	0.4901	0.9728	1.0000	0.9531	**0.0000**	0.0558	**0.0000**
DFO_SU	DFO_FA	0.0722	0.0900	**0.0115**	0.0667	**0.0000**	**0.0031**	**0.0000**
DFO_SU	DFO_SW	**0.0053**	**0.0108**	**0.0131**	**0.0149**	**0.0005**	**0.0002**	0.0759
DFO_SU	BFO_HA	0.6559	0.3705	0.9962	0.5622	**0.0000**	0.0646	**0.0000**
DFO_SU	BFO_CL	0.3593	0.3463	0.1599	0.6404	**0.0000**	0.7510	**0.0000**
DFO_SU	BFO_TU	0.8081	0.0701	0.2850	0.2083	**0.0453**	0.6706	**0.0004**
DFO_SU	BFO_JA	0.0978	0.0892	0.2171	0.0838	**0.0000**	0.6200	**0.0001**
DFO_SU	BFO_FR	**0.0303**	0.1785	**0.0084**	0.3835	**0.0001**	**0.0002**	**0.0000**
DFO_SU	BFO_EL	0.4325	0.0874	**0.0006**	0.0594	**0.0000**	0.7412	**0.0001**
DFO_SU	BFO_BR	**0.0087**	**0.0026**	**0.0070**	0.4262	**0.0037**	**0.0001**	**0.0004**
DFO_SU	AFO_MN	0.1097	0.2752	0.3327	0.7949	**0.0109**	**0.0073**	**0.0000**
DFO_SU	AFO_BA	0.6066	0.5587	0.8254	0.6463	**0.0005**	0.6756	**0.0000**
DFO_SU	AFO_JA	0.3734	0.8786	0.3586	0.5100	**0.0000**	**0.0265**	**0.0000**
DFO_SU	AFO_CA	0.1553	0.5866	0.2438	0.1065	**0.0000**	**0.0394**	**0.0001**
DFO_SU	AFO_BR	0.8828	0.9243	0.9981	0.3715	**0.0000**	0.3940	**0.0000**
DFO_SU	AFO_HR	**0.0204**	0.1071	0.0956	0.0558	**0.0000**	**0.0024**	0.1297
DFO_BR	DFO_HS	0.2090	0.0881	0.1724	0.0688	**0.0000**	0.4675	**0.0015**
DFO_BR	DFO_TP	0.8169	0.1972	0.5300	0.5001	0.7440	0.8298	**0.0014**

Table 10 (continued)

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
DFO_BR	DFO_MA	0.5071	0.5528	0.7177	0.9787	0.7058	0.1112	**0.0000**
DFO_BR	DFO_FA	**0.0200**	0.3037	**0.0459**	0.1492	0.9137	**0.0041**	**0.0000**
DFO_BR	DFO_SW	**0.0026**	0.2297	0.0998	**0.0422**	0.2159	**0.0001**	0.9020
DFO_BR	BFO_HA	0.4656	0.7679	0.7479	0.5418	0.8558	**0.0289**	0.0508
DFO_BR	BFO_CL	0.3459	0.9916	0.3215	0.7511	**0.0052**	0.7204	**0.0197**
DFO_BR	BFO_TU	0.5545	0.8919	0.5198	0.1829	**0.0044**	0.3475	0.1252
DFO_BR	BFO_JA	0.4828	**0.0476**	0.1381	0.0513	0.4381	0.3644	**0.0053**
DFO_BR	BFO_FR	**0.0084**	0.7158	0.0513	0.6100	0.3346	**0.0000**	**0.0002**
DFO_BR	BFO_EL	0.6043	**0.0065**	**0.0002**	0.1432	0.2052	0.5916	**0.0328**
DFO_BR	BFO_BR	**0.0044**	**0.0259**	**0.0490**	0.3277	**0.0139**	**0.0000**	0.3269
DFO_BR	AFO_MN	**0.0388**	0.9962	0.7226	0.9611	**0.0382**	**0.0109**	**0.0002**
DFO_BR	AFO_BA	0.8623	0.1492	0.6974	0.2403	0.2257	0.1937	**0.0079**
DFO_BR	AFO_JA	0.2056	0.4909	0.6834	0.9972	0.2472	**0.0295**	**0.0155**
DFO_BR	AFO_CA	0.1581	0.5009	0.2471	0.1940	0.4402	**0.0191**	**0.0192**
DFO_BR	AFO_BR	0.6800	0.2301	0.5683	0.2621	0.1601	0.1150	**0.0125**
DFO_BR	AFO_HR	**0.0060**	0.5641	0.4492	0.3752	0.1139	**0.0014**	0.7200
DFO_HS	DFO_TP	0.1078	0.6612	0.2457	0.1772	**0.0013**	0.3081	0.6320
DFO_HS	DFO_MA	**0.0130**	0.1519	0.0975	**0.0234**	**0.0000**	**0.0023**	**0.0195**
DFO_HS	DFO_FA	**0.0012**	**0.0035**	**0.0007**	**0.0011**	**0.0003**	**0.0002**	**0.0305**
DFO_HS	DFO_SW	**0.0000**	**0.0018**	**0.0007**	**0.0003**	**0.0000**	**0.0000**	**0.0021**
DFO_HS	BFO_HA	**0.0154**	**0.0241**	0.2125	0.2499	**0.0003**	**0.0003**	0.2199
DFO_HS	BFO_CL	**0.0364**	0.0669	**0.0171**	**0.0415**	0.4888	0.1183	0.5339
DFO_HS	BFO_TU	0.2220	**0.0088**	**0.0149**	0.9008	**0.0000**	0.1014	0.0638
DFO_HS	BFO_JA	0.7273	0.6789	0.3359	0.5109	**0.0023**	**0.0214**	0.5457
DFO_HS	BFO_FR	**0.0002**	0.1084	**0.0002**	**0.0140**	**0.0000**	**0.0000**	0.7312
DFO_HS	BFO_EL	0.3642	0.1080	**0.0230**	0.9587	**0.0040**	0.1803	0.3194
DFO_HS	BFO_BR	**0.0001**	**0.0001**	**0.0004**	**0.0284**	**0.0000**	**0.0000**	**0.0242**
DFO_HS	AFO_MN	**0.0009**	**0.0415**	**0.0081**	**0.0454**	**0.0000**	**0.0006**	0.6825
DFO_HS	AFO_BA	0.2733	0.8105	0.3032	0.2474	**0.0000**	**0.0302**	0.5620
DFO_HS	AFO_JA	**0.0077**	0.3384	**0.0373**	**0.0377**	**0.0016**	**0.0014**	0.1957
DFO_HS	AFO_CA	**0.0037**	0.1900	**0.0045**	**0.0009**	**0.0000**	**0.0010**	0.5545
DFO_HS	AFO_BR	**0.0301**	0.4695	0.1611	0.2695	**0.0100**	**0.0173**	0.3772
DFO_HS	AFO_HR	**0.0000**	**0.0133**	**0.0031**	**0.0007**	**0.0061**	**0.0000**	**0.0005**
DFO_TP	DFO_MA	0.5282	0.8267	0.8482	0.6594	0.6172	0.1244	**0.0109**
DFO_TP	DFO_FA	**0.0413**	**0.0342**	**0.0179**	**0.0225**	0.4673	**0.0035**	**0.0372**
DFO_TP	DFO_SW	**0.0044**	**0.0043**	**0.0061**	**0.0165**	0.0508	**0.0003**	**0.0027**
DFO_TP	BFO_HA	0.6377	0.5470	0.9914	0.9318	0.8010	**0.0215**	0.2455
DFO_TP	BFO_CL	0.4996	0.3082	0.2014	0.5846	**0.0062**	0.9801	0.4365
DFO_TP	BFO_TU	0.8771	0.1602	0.4017	0.3396	**0.0016**	0.8503	0.0524
DFO_TP	BFO_JA	0.1010	0.4174	0.5924	**0.0436**	0.8185	0.4994	0.7421
DFO_TP	BFO_FR	**0.0198**	0.1772	**0.0129**	0.3053	0.2125	**0.0001**	1.0000
DFO_TP	BFO_EL	0.3495	0.1301	**0.0016**	0.2353	0.2303	0.9606	0.2568
DFO_TP	BFO_BR	**0.0066**	**0.0026**	**0.0353**	0.3586	**0.0266**	**0.0001**	**0.0308**
DFO_TP	AFO_MN	**0.0274**	0.3467	0.2238	0.6182	**0.0087**	**0.0100**	0.5482

Table 10 (continued)

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
DFO_TP	AFO_BA	0.6789	0.5314	0.6993	0.5771	0.0732	0.3648	0.4339
DFO_TP	AFO_JA	0.1849	0.8712	0.3698	0.2678	0.5901	**0.0289**	0.2762
DFO_TP	AFO_CA	0.1952	0.5547	0.1233	**0.0322**	0.2899	**0.0181**	0.5488
DFO_TP	AFO_BR	0.9109	0.8539	0.7699	0.7729	0.3994	0.1853	0.3049
DFO_TP	AFO_HR	**0.0261**	0.1265	0.1232	0.0556	0.1927	**0.0006**	**0.0005**
DFO_MA	DFO_FA	0.1921	0.1360	0.0616	**0.0488**	0.7517	0.2468	0.5774
DFO_MA	DFO_SW	**0.0236**	**0.0245**	**0.0111**	0.0619	0.6798	**0.0139**	**0.0000**
DFO_MA	BFO_HA	0.8999	0.6959	0.8416	0.2930	0.9485	0.7714	**0.0008**
DFO_MA	BFO_CL	0.8740	0.2484	0.2792	0.9805	**0.0003**	0.1910	**0.0074**
DFO_MA	BFO_TU	0.5488	0.1190	0.4285	0.0765	**0.0204**	0.4930	**0.0000**
DFO_MA	BFO_JA	**0.0266**	0.2357	0.1350	**0.0068**	0.1786	0.3497	**0.0012**
DFO_MA	BFO_FR	0.0859	0.3307	0.0518	0.7903	0.6589	**0.0289**	**0.0181**
DFO_MA	BFO_EL	0.0529	0.0638	**0.0003**	0.0752	**0.0463**	0.1110	**0.0009**
DFO_MA	BFO_BR	**0.0459**	**0.0032**	**0.0159**	0.9157	0.0823	**0.0143**	**0.0000**
DFO_MA	AFO_MN	0.2381	0.4675	0.5766	0.9363	0.0521	0.3679	**0.0069**
DFO_MA	AFO_BA	0.3155	0.3827	0.6660	0.3115	0.2556	0.3231	**0.0022**
DFO_MA	AFO_JA	0.5457	0.7916	0.6513	0.9688	0.2585	0.6927	**0.0003**
DFO_MA	AFO_CA	0.3954	0.8679	0.2591	0.1620	0.7440	0.5770	**0.0013**
DFO_MA	AFO_BR	0.4481	0.5917	0.7516	0.2225	0.0570	0.7426	**0.0002**
DFO_MA	AFO_HR	**0.0367**	0.2188	0.1363	0.2037	**0.0181**	0.0762	**0.0000**
DFO_FA	DFO_SW	0.2637	0.6313	0.7257	0.6190	0.2953	0.0938	**0.0000**
DFO_FA	BFO_HA	0.1318	0.1639	**0.0179**	**0.0187**	0.8462	0.2728	**0.0013**
DFO_FA	BFO_CL	0.2482	0.5497	0.4262	0.1196	**0.0047**	**0.0197**	**0.0180**
DFO_FA	BFO_TU	0.0815	0.6297	0.2328	**0.0097**	**0.0062**	0.0658	**0.0002**
DFO_FA	BFO_JA	**0.0014**	**0.0050**	**0.0018**	**0.0003**	0.4487	**0.0229**	**0.0053**
DFO_FA	BFO_FR	0.6141	0.3764	0.8169	0.4101	0.2486	0.1992	0.0989
DFO_FA	BFO_EL	**0.0057**	**0.0004**	**0.0000**	**0.0016**	0.1987	**0.0051**	**0.0015**
DFO_FA	BFO_BR	0.4763	0.3574	0.8937	0.3082	0.0826	**0.0324**	**0.0000**
DFO_FA	AFO_MN	0.6886	0.2519	0.2081	0.0883	**0.0281**	0.6791	0.0717
DFO_FA	AFO_BA	0.0669	**0.0449**	**0.0310**	**0.0258**	0.2153	**0.0418**	**0.0085**
DFO_FA	AFO_JA	0.4490	0.0782	0.0994	0.1043	0.3014	0.5205	**0.0002**
DFO_FA	AFO_CA	0.6758	0.1053	0.3230	0.6402	0.5172	0.5270	**0.0007**
DFO_FA	AFO_BR	**0.0500**	**0.0220**	**0.0126**	**0.0065**	0.0917	0.1028	**0.0022**
DFO_FA	AFO_HR	0.6789	0.7501	0.3642	0.7147	0.1332	0.6252	**0.0000**
DFO_SW	BFO_HA	**0.0123**	0.0785	**0.0174**	**0.0026**	0.1515	**0.0475**	0.0593
DFO_SW	BFO_CL	**0.0121**	0.2360	0.3707	0.1099	**0.0001**	**0.0004**	**0.0083**
DFO_SW	BFO_TU	**0.0011**	0.4050	0.2033	**0.0033**	**0.0310**	**0.0007**	0.0976
DFO_SW	BFO_JA	**0.0003**	**0.0015**	**0.0031**	**0.0000**	**0.0496**	**0.0004**	**0.0171**
DFO_SW	BFO_FR	0.5051	0.1165	0.8427	0.1131	0.8340	0.8487	**0.0001**
DFO_SW	BFO_EL	**0.0006**	**0.0000**	**0.0000**	**0.0005**	**0.0224**	**0.0001**	**0.0456**
DFO_SW	BFO_BR	0.7572	0.4820	0.9425	0.3036	0.4167	0.3530	0.2251
DFO_SW	AFO_MN	0.1477	0.0585	0.1678	0.0654	0.3487	0.1192	**0.0003**
DFO_SW	AFO_BA	**0.0019**	**0.0120**	**0.0093**	**0.0132**	0.8132	**0.0021**	**0.0075**
DFO_SW	AFO_JA	**0.0226**	**0.0224**	**0.0490**	**0.0481**	**0.0276**	0.0766	**0.0195**

Table 10 (continued)

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
DFO_SW	AFO_CA	0.1274	**0.0378**	0.0894	0.3498	0.4823	0.0595	**0.0070**
DFO_SW	AFO_BR	**0.0021**	**0.0033**	**0.0049**	**0.0018**	**0.0089**	**0.0054**	**0.0209**
DFO_SW	AFO_HR	0.3880	0.2626	0.2537	0.4828	**0.0083**	0.5035	0.4915
BFO_HA	BFO_CL	0.9513	0.9731	0.2271	0.3662	**0.0006**	0.0777	0.3749
BFO_HA	BFO_TU	0.5064	0.3666	0.5149	0.4142	**0.0010**	0.1618	0.6592
BFO_HA	BFO_JA	**0.0167**	0.0676	0.5356	0.0812	0.3518	0.1417	0.4761
BFO_HA	BFO_FR	**0.0224**	0.8104	**0.0237**	0.2662	0.2099	**0.0471**	0.0816
BFO_HA	BFO_EL	0.2871	**0.0028**	**0.0001**	0.1861	0.1862	**0.0292**	0.7619
BFO_HA	BFO_BR	**0.0058**	**0.0055**	**0.0134**	0.1719	**0.0450**	**0.0185**	0.3954
BFO_HA	AFO_MN	0.1612	0.9330	0.2395	0.2790	**0.0200**	0.6211	0.1104
BFO_HA	AFO_BA	0.3801	0.1175	0.7059	0.8151	0.3973	0.1023	0.6299
BFO_HA	AFO_JA	0.2851	0.5953	0.6602	0.1439	0.2956	0.9352	0.6379
BFO_HA	AFO_CA	0.3445	0.7335	0.2003	0.0535	0.4257	0.7987	0.7599
BFO_HA	AFO_BR	0.6436	0.1962	0.9308	0.7785	0.1203	0.4216	0.5435
BFO_HA	AFO_HR	**0.0362**	0.6174	0.1711	**0.0156**	0.2814	0.1582	0.0599
BFO_CL	BFO_TU	0.5961	0.6283	0.5619	0.1265	**0.0000**	0.8120	0.0929
BFO_CL	BFO_JA	0.1294	0.1236	**0.0498**	**0.0182**	**0.0129**	0.7315	0.8010
BFO_CL	BFO_FR	0.0848	0.9880	0.3444	0.9082	**0.0001**	**0.0008**	0.4897
BFO_CL	BFO_EL	0.1677	**0.0034**	**0.0001**	0.0667	0.0571	0.7023	0.4701
BFO_CL	BFO_BR	**0.0489**	0.0708	0.3673	0.6235	**0.0000**	**0.0004**	**0.0337**
BFO_CL	AFO_MN	0.3269	0.9241	0.7662	0.7581	**0.0000**	0.0631	0.4395
BFO_CL	AFO_BA	0.3296	0.1964	0.3054	0.4243	**0.0000**	0.7863	0.7680
BFO_CL	AFO_JA	0.4911	0.3936	0.5403	0.8276	**0.0135**	0.1133	0.4945
BFO_CL	AFO_CA	0.4133	0.5021	0.7952	0.5235	**0.0002**	0.0790	0.7057
BFO_CL	AFO_BR	0.7020	0.2264	0.1506	0.1050	**0.0315**	0.3070	0.6009
BFO_CL	AFO_HR	0.0910	0.5960	0.9389	0.4850	0.0681	**0.0104**	**0.0010**
BFO_TU	BFO_JA	0.2404	**0.0053**	0.0810	0.7371	**0.0001**	0.9213	0.2781
BFO_TU	BFO_FR	**0.0052**	0.4615	0.1048	0.0762	**0.0171**	**0.0018**	**0.0107**
BFO_TU	BFO_EL	0.3277	**0.0009**	**0.0001**	0.6593	**0.0000**	0.5430	0.3569
BFO_TU	BFO_BR	**0.0065**	0.2628	0.2252	0.0810	0.2388	**0.0005**	0.7083
BFO_TU	AFO_MN	**0.0244**	0.2456	0.7230	0.2342	0.3132	**0.0477**	**0.0260**
BFO_TU	AFO_BA	0.7812	0.0593	0.3814	0.7851	0.0869	0.7293	0.0934
BFO_TU	AFO_JA	0.1048	0.2020	0.7954	0.1180	**0.0003**	0.2247	0.4693
BFO_TU	AFO_CA	0.2572	0.2207	0.5960	**0.0134**	**0.0051**	0.2893	0.2474
BFO_TU	AFO_BR	0.5742	0.0821	0.2215	0.4847	**0.0001**	0.5475	0.3110
BFO_TU	AFO_HR	**0.0132**	0.7617	0.4821	**0.0098**	**0.0002**	**0.0230**	0.0841
BFO_JA	BFO_FR	**0.0000**	**0.0273**	**0.0001**	**0.0168**	0.0607	**0.0003**	0.2326
BFO_JA	BFO_EL	0.7839	0.1532	**0.0031**	0.6756	0.6665	0.2554	0.5177
BFO_JA	BFO_BR	**0.0001**	**0.0001**	**0.0009**	**0.0069**	**0.0022**	**0.0000**	0.0929
BFO_JA	AFO_MN	**0.0051**	**0.0174**	**0.0225**	**0.0129**	**0.0030**	**0.0447**	0.2076
BFO_JA	AFO_BA	0.6808	0.8758	0.5786	0.3059	**0.0334**	0.9008	0.8212
BFO_JA	AFO_JA	**0.0351**	0.3617	0.0845	**0.0178**	0.9625	0.1615	0.6546
BFO_JA	AFO_CA	**0.0049**	0.3464	**0.0100**	**0.0000**	0.1491	0.0760	0.9725
BFO_JA	AFO_BR	0.1825	0.7925	0.6562	0.1436	0.2863	0.6541	0.9066

Table 10 (continued)

Scenario1	Scenario2	everyone	first_person	middle_ person	last_person	cutter	univers.	likelihood
BFO_JA	AFO_HR	**0.0004**	**0.0369**	**0.0121**	**0.0001**	0.2403	**0.0116**	**0.0014**
BFO_FR	BFO_EL	**0.0007**	**0.0024**	**0.0000**	**0.0430**	**0.0243**	**0.0001**	0.1026
BFO_FR	BFO_BR	0.7396	0.0562	0.9728	0.8804	0.1931	0.6061	**0.0028**
BFO_FR	AFO_MN	0.3402	0.6681	0.0838	0.6928	0.3363	0.1337	0.9534
BFO_FR	AFO_BA	**0.0126**	0.1318	**0.0079**	0.1637	0.8240	**0.0013**	0.3970
BFO_FR	AFO_JA	0.2461	0.1639	**0.0327**	0.7905	0.0506	**0.0260**	0.1602
BFO_FR	AFO_CA	0.1324	0.2862	0.1430	0.7652	0.6541	0.0653	0.3283
BFO_FR	AFO_BR	**0.0162**	0.2348	**0.0047**	0.1717	**0.0226**	**0.0027**	0.1385
BFO_FR	AFO_HR	0.8957	0.6338	0.4099	0.6166	**0.0155**	0.7785	**0.0001**
BFO_EL	BFO_BR	**0.0002**	**0.0000**	**0.0000**	**0.0345**	**0.0002**	**0.0000**	0.2260
BFO_EL	AFO_MN	**0.0089**	**0.0052**	**0.0000**	**0.0447**	**0.0011**	**0.0072**	0.2265
BFO_EL	AFO_BA	0.5790	0.0911	**0.0007**	0.2429	**0.0081**	0.5451	0.7783
BFO_EL	AFO_JA	**0.0314**	0.0710	**0.0004**	0.0732	0.3790	**0.0230**	0.8962
BFO_EL	AFO_CA	**0.0255**	**0.0345**	**0.0000**	**0.0043**	**0.0191**	**0.0171**	0.7502
BFO_EL	AFO_BR	0.2964	0.1091	**0.0012**	0.3388	0.5868	0.3081	0.9016
BFO_EL	AFO_HR	**0.0059**	**0.0015**	**0.0000**	**0.0008**	0.9453	**0.0018**	**0.0038**
BFO_BR	AFO_MN	0.1362	0.0636	0.1382	0.8231	0.8852	**0.0465**	**0.0013**
BFO_BR	AFO_BA	**0.0080**	**0.0013**	**0.0367**	0.2202	0.3267	**0.0017**	0.2366
BFO_BR	AFO_JA	0.1207	**0.0005**	0.0533	0.4999	**0.0005**	**0.0140**	0.2834
BFO_BR	AFO_CA	0.2180	**0.0028**	0.1815	0.5277	0.1003	**0.0143**	0.1146
BFO_BR	AFO_BR	**0.0054**	**0.0005**	**0.0088**	0.1892	**0.0003**	**0.0023**	0.1244
BFO_BR	AFO_HR	0.8312	0.0631	0.4024	0.7129	**0.0006**	0.5031	0.1892
AFO_MN	AFO_BA	**0.0373**	0.1171	0.2233	0.3143	0.2948	0.0933	0.2197
AFO_MN	AFO_JA	0.4770	0.2520	0.4602	0.9951	**0.0019**	0.3867	0.1078
AFO_MN	AFO_CA	0.6662	0.3075	0.7797	0.2243	0.0789	0.7451	0.3375
AFO_MN	AFO_BR	0.0855	0.1120	0.1801	0.7379	**0.0004**	0.1670	0.0964
AFO_MN	AFO_HR	0.5652	0.5427	0.5334	0.3501	**0.0007**	0.3941	**0.0001**
AFO_BA	AFO_JA	0.2087	0.6851	0.4934	0.1770	0.0507	0.2085	0.4477
AFO_BA	AFO_CA	0.0610	0.4705	0.1079	**0.0151**	0.4383	0.0909	0.9597
AFO_BA	AFO_BR	0.6675	0.9350	0.9289	0.8694	**0.0200**	0.5074	0.7100
AFO_BA	AFO_HR	**0.0035**	0.0610	0.0722	**0.0110**	**0.0037**	**0.0246**	**0.0025**
AFO_JA	AFO_CA	0.9262	0.7176	0.3245	0.2117	0.1669	0.7737	0.7711
AFO_JA	AFO_BR	0.4600	0.3814	0.7048	0.3363	0.6374	0.2145	0.5911
AFO_JA	AFO_HR	0.2319	0.1083	0.4105	0.2188	0.5315	0.1147	**0.0131**
AFO_CA	AFO_BR	0.2065	0.4646	0.1812	**0.0153**	**0.0282**	0.2662	0.7071
AFO_CA	AFO_HR	0.2848	0.1267	0.9936	0.9962	**0.0449**	0.3155	**0.0077**
AFO_BR	AFO_HR	**0.0258**	0.0622	0.0782	**0.0312**	0.9389	**0.0195**	**0.0071**
Perc. of affected pair		37.67%	27.67%	35.33%	31.00%	46.00%	42.33%	58.67%

The null-hypothesis is rejected if $p < 0.05$, these cases are reported in bold. p-values of EVs for any pair of scenarios

Author Contributions All authors contributed equally.

Funding Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE Agreement.

Data availability All the data collected are available on https://github.com/aloreggia/SEP-net/data.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval The study has been approved by the Massachusetts Institute of Technology Institutional Review Board (Protocol 0812003014). The review process ensures compliance with university-mandated ethical guidelines for research conducted with human subjects. All subjects provided informed consent prior to participating. See https://couhes.mit.edu for details of the review process.

Consent to participate Not applicable.

Consent to publish Not applicable.

References

1. Russell, S., Hauert, S., Altman, R., & Veloso, M. (2015). Ethics of artificial intelligence. *Nature, 521*(7553), 415–416.
2. O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
3. Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
4. Rossi, F., & Mattei, N. (2019). Building ethically bounded AI. In: Proc. of the 33rd AAAI(Blue Sky Track).
5. Amodei, D., Olah, C., Steinhardt, J., Christiano, P.F., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565
6. Hart, H. (1958). Positivism and the separation of law and morals. *Harvard Law Review, 71*, 593–607.
7. Clark, J., & Amodei, D. (2016). Faulty reward functions in the wild. Retrieved 1 Aug 2023 from https://blog.openai.com/faulty-reward-functions
8. Branwen, G. (2023). The Neural Net Tank Urban Legend. Retrieved 1 Aug 2023 from https://gwern.net/tank#alternative-examples
9. ACM US Public Policy Working Group: Statement on algorithmic transparency and accountability. Retrieved 1 Aug 2023 from https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
10. National Institute of Standards and Technology (NIST): AI Risk Management Framework: Second Draft. Retrieved 1 Aug 2023 from https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf
11. Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

12. Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine, 36*(4), 105–114.
13. Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.
14. Balakrishnan, A., Bouneffouf, D., Mattei, N., & Rossi, F. (2019). Incorporating behavioral constraints in online AI systems. In *Proc. of the 33rd AAAI*.
15. Loreggia, A., Mattei, N., Rahgooy, T., Rossi, F., Srivastava, B., & Venable, K.B. (2022). Making human-like moral decisions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22, pp. 447–454. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3514094.3534174
16. Svegliato, J., Nashed, S.B., & Zilberstein, S. (2021). Ethically compliant sequential decision making. In *Proceedings of the 35th AAAI International Conference on Artificial Intelligence (AAAI)*.
17. Wallach, W., Allen, C., & Smit, I. (2008). *Machine morality: bottom-up and top-down approaches for modelling human moral faculties.,22*, 565–582.
18. Loreggia, A., Mattei, N., Rossi, F., & Venable, K.B. (2018). Preferences and ethical principles in decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18, p. 222. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3278721.3278723
19. Hansson, S.O. (2001). The Structure of Values and Norms. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press.
20. Boutilier, C., Brafman, R., Domshlak, C., Hoos, H. H., & Poole, D. (2004). CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research, 21*, 135–191.
21. Alashaikh, A., & Alanazi, E. (2021). Conditional preference networks for cloud service selection and ranking with many irrelevant attributes. *IEEE Access, 9*, 131214–131222.
22. Mohajeriparizi, M., Sileno, G., & Engers, T. (2022). Preference-based goal refinement in bdi agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 917–925.
23. Cornelio, C., Goldsmith, J., Grandi, U., Mattei, N., Rossi, F., & Venable, K. B. (2021). Reasoning with pcp-nets. *Journal of Artificial Intelligence Research, 72*, 1103–1161.
24. Kahneman, D. (2011). *Thinking*. Straus and Giroux, New York: Fast and Slow. Farrar.
25. Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3.
26. Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.
27. Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.
28. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59.
29. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.
30. Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567.
31. Doris, J.M., Group, M.P.R., et al. (2010). The moral psychology handbook. OUP Oxford.
32. Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002.
33. Knobe, J. (2007). Experimental philosophy. *Philosophy Compass, 2*(1), 81–92.
34. Alexander, J. (2012). *Experimental philosophy: An introduction*. Polity Press.
35. Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A. C., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., & Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences, 26*(5), 388–405. https://doi.org/10.1016/j.tics.2022.02.009
36. Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences, 117*(42), 26158–26169.
37. Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition, 167*, 107–123.
38. Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J.B., & Rahwan, I. (2018). A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 197–203.
39. Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications, 10*(1), 1–14.

40. Engelmann, N., & Waldmann, M. R. (2022). How to weigh lives. a computational model of moral judgment in multiple-outcome structures. *Cognition, 218*, 104910.
41. Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2021). Can Machines Learn Morality? The Delphi Experiment. arXiv. https://doi.org/10.48550/ARXIV.2110.07574 . arXiv:2110.07574
42. Awad, E., Anderson, M., Anderson, S. L., & Liao, B. (2020). An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence, 287*, 103349.
43. Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2020). *The moral psychology of AI and the ethical opt-out problem*. Oxford University Press.
44. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.
45. Theodorou, A., Wortham, R.H., & Bryson, J.J. (2016). 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.
46. Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.
47. Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A.D. (2017). A voting-based system for ethical decision making. In: Proc. of the 32nd AAAI.
48. Iacca, G., Lagioia, F., Loreggia, A., & Sartor, G. (2020). A genetic approach to the ethical knob. In: Legal Knowledge and Information Systems: JURIX 2020, vol. 334, pp. 103–112. IOS Press, Amsterdam. https://doi.org/10.3233/FAIA200854
49. Grandi, U., Loreggia, A., Rossi, F., & Saraswat, V.A. (2014). From sentiment analysis to preference aggregation. In *International Symposium on Artificial Intelligence and Mathematics*, ISAIM 2014, Fort Lauderdale, FL, USA, January 6–8, 2014.
50. Loreggia, A., Mattei, N., Rossi, F., & Venable, K.B. (2019). Metric learning for value alignment. In *AISafety@IJCAI. CEUR Workshop Proceedings*, vol. 2419. CEUR-WS.org, Aachen.
51. Domshlak, C., Hüllermeier, E., Kaci, S., & Prade, H. (2011). Preferences in AI: An overview. *Artificial Intelligence, 175*(7), 1037–1052.
52. Rossi, F., & Loreggia, A. (2019). Preferences and ethical priorities: Thinking fast and slow in AI. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*. AAMAS '19, pp. 3–4. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.
53. Sen, A. (1974). Choice, ordering and morality. In S. Körner (Ed.), *Practical Reason.* Blackwell.
54. Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social Research, 44*(4), 623.
55. Loreggia, A., Lorini, E., & Sartor, G. (2022). Modelling ceteris paribus preferences with deontic logic. *Journal of Logic and Computation, 32*(2), 347–368. https://doi.org/10.1093/logcom/exab088
56. Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence, 283*, 103261.
57. Lee, M.K., Kusbit, D., Kahng, A., Kim, J.T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., et al. (2019). Webuildai: Participatory framework for algorithmic governance. In *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW), 1–35.
58. Rossi, F., Venable, K.B., & Walsh, T. (2011). A short introduction to preferences: Between artificial intelligence and social choice, pp. 1–102. Morgan and Claypool, San Rafael, California (USA).
59. Brandt, F., Conitzer, V., Endriss, U., Lang, J., Procaccia, A.D. (eds.): Handbook of Computational Social Choice. Cambridge University Press, Pennsylvania (2016). http://dblp.uni-trier.de/db/reference/choice/choice2016.html
60. Wang, H., Shao, S., Zhou, X., Wan, C., & Bouguettaya, A. (2009). Web service selection with incomplete or inconsistent user preferences. In *Proc. 7th International Conference on Service-Oriented Computing*, pp. 83–98. Springer, Berlin, Heidelberg.
61. Pu, P., Faltings, B., Chen, L., Zhang, J., & Viappiani, P. (2011). Usability guidelines for product recommenders based on example critiquing research. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 511–545). Springer.
62. Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V.R., & Yang, Q. (2018). Building ethics into artificial intelligence. In: Proc. 27th IJCAI, pp. 5527–5533.
63. Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K., Campbell, M., Singh, M., & Rossi, F. (2019). Teaching AI agents ethical values using reinforcement learning and policy orchestration. In: Proc. of the 28th IJCAI.
64. Alkoby, S., Rath, A., & Stone, P. (2019). Teaching social behavior through human reinforcement for ad hoc teamwork-the STAR framework. In: Proc. of The 18th AAMAS.

65. Arnold, T., Thomas, Kasenberg, D., & Scheutzs, M. (2017). Value alignment or misalignment - what will keep systems accountable? In: AI, Ethics, and Society, Papers from the 2017 AAAI Workshop.
66. Loreggia, A., Mattei, N., Rossi, F., & Venable, K.B. (2020). Modeling and reasoning with preferences and ethical priorities in AI systems. In: Liao, S.M. (ed.) Ethics of Artificial Intelligence, New York, pp. 127–154. Chap. 4. https://doi.org/10.1093/oso/9780190905033.003.0005
67. Loreggia, A., Calegari, R., Lorini, E., Rossi, F., & Sartor, G. (2022). How to model contrary-to-duty with gcp-nets. *Intelligenza Artificiale, 16*(2), 185–198.
68. Loreggia, A., Mattei, N., Rossi, F., & Venable, K.B. (2020). CPMetric: Deep siamese networks for metric learning on structured preferences. In: El Fallah Seghrouchni, A., Sarne, D. (eds.) Artificial Intelligence. IJCAI 2019 International Workshops, pp. 217–234. Springer, Cham.
69. Loreggia, A., Mattei, N., Rossi, F., & Venable, K.B. (2018). On the distance between cp-nets. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. AAMAS '18, pp. 955–963. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.
70. Jiang, L., Hwang, J.D., Bhagavatula, C., Bras, R.L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., et al. (2021). Can machines learn morality? the delphi experiment. arXiv preprint arXiv:2110.07574
71. Lieder, F., & Griffiths, T.L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*. **43**.
72. Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.
73. Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129.
74. Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J.B. (2015). Inference of intention and permissibility in moral decision making. In: CogSci. Citeseer.
75. Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin, 142*(11), 1179.
76. Gauthier, D. (1986). *Morals by Agreement*. Oxford University Press on Demand.
77. Rawls, J. (1971). *A theory of justice*. Harvard University Press.
78. Scanlon, T., et al. (1998). *What we owe to each other*. Harvard University Press.
79. Habermas, J. (1990). *Moral consciousness and communicative action*. MIT press.
80. Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J.B. (2022). When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment.
81. Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(1), 59–78.
82. André, J.-B., Debove, S., Fitouchi, L., & Baumard, N. (2022). Moral cognition as a Nash product maximizer: An evolutionary contractualist account of morality. PsyArXiv. https://doi.org/10.31234/osf.io/2hxgu
83. Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772.
84. Kant, I., & Schneewind, J. B. (2002). *Groundwork for the Metaphysics of Morals*. Yale University Press.
85. Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023). Resource-rational contractualism: A triple theory of moral cognition.
86. Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford University Press.
87. Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.
88. Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F., & Tenenbaum, J.B. (2018). The cognitive mechanisms of contractualist moral decision-making. In: CogSci. Citeseer.
89. Stich, S. (2018). The quest for the boundaries of morality. In: The Routledge handbook of moral epistemology, pp. 15–37. Routledge, New York.
90. Levine, S., Rottman, J., Davis, T., O'Neill, E., Stich, S., & Machery, E. (2021). Religious affiliation and conceptions of the moral domain. *Social Cognition, 39*(1), 139–165.
91. Kwon, J., Tenenbaum, J., & Levine, S. (2022). Flexibility in moral cognition: When is it okay to break the rules? In *Proceedings of the 44th annual conference of the cognitive science society*.
92. Kwon, J., Zhi-Xuan, T., Tenenbaum, J., & Levine, S. When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments.
93. Allen, T.E. (2013). CP-nets with indifference. In: 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1488–1495. IEEE.

94. Goldsmith, J., Lang, J., Truszczyński, M., & Wilson, N. (2008). The computational complexity of dominance and consistency in CP-nets. *Journal of Artificial Intelligence Research, 33*(1), 403–432.
95. Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., Loreggia, A., Murgesan, K., Mattei, N., Rossi, F., et al. (2021). Thinking fast and slow in AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 15042–15046.
96. Difallah, D., Filatova, E., & Ipeirotis, P. (2018) Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 135–143.
97. Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*(9), 536–554.
98. Mann, H., Whitney, D., et al. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50–60.
99. Cornelio, C., Goldsmith, J., Mattei, N., Rossi, F., & Venable, K.B. (2013). Updates and uncertainty in CP-nets. In: Proc. of the 26th AUSAI.
100. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378.
101. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics, 29*, 1189–1232.
102. Cornelio, C., Donini, M., Loreggia, A., Pini, M. S., & Rossi, F. (2021). Voting with random classifiers (VORACE): Theoretical and experimental analysis. *Autonomous Agents and Multi-Agent Systems, 35*(2), 22. https://doi.org/10.1007/s10458-021-09504-y
103. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
104. Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*(3), 530–542.
105. Levine, S., & Leslie, A. (2021). Preschoolers use the means principle to make moral judgments.
106. Parfit, D. (2011). *On what matters* (Vol. 1). Oxford University Press.
107. Azari Soufiani, H., Diao, H., Lai, Z., & Parkes, D.C. (2013). Generalized random utility models with multiple types. *Advances in Neural Information Processing Systems*. **26**.
108. Brafman, R.I., & Chernyavsky, Y. (2005). Planning with goal preferences and constraints. In: ICAPS, pp. 182–191.
109. Benton, J., Coles, A., & Coles, A. (2012). Temporal planning with preferences and time-dependent continuous costs. In: Proc. 22nd ICAPS.
110. Gerevini, A., & Long, D. (2005). Plan constraints and preferences in pddl3. In: Technical Report 2005-08-07, Department of Electronics for Automation.
111. Pallagani, V., Muppasani, B., Srivastava, B., Rossi, F., Horesh, L., Murugesan, K., Loreggia, A., Fabiano, F., Joseph, R., Kethepalli, Y., et al. (2023). Plansformer tool: demonstrating generation of symbolic plans using transformers. In: IJCAI, vol. 2023, pp. 7158–7162. In *International Joint Conferences on Artificial Intelligence*.

Authors and Affiliations

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

✉ Andrea Loreggia
andrea.loreggia@gmail.com

Edmond Awad
e.awad@exeter.ac.uk

Sydney Levine
smlevine@mit.edu

Nicholas Mattei
nsmattei@tulane.edu

Iyad Rahwan
sekrahwan@mpib-berlin.mpg.de

Francesca Rossi
francesca.rossi2@ibm.com

Kartik Talamadupula
kartik.t@symbl.ai

Joshua Tenenbaum
jbt@mit.edu

Max Kleiman-Weiner
maxkw@uw.edu

[1] University of Exeter, Exeter, UK

[2] Massachusetts Institute of Technology, Cambridge, USA

[3] Harvard University, Cambridge, USA

[4] University of Brescia, Brescia, Italy

[5] Tulane University, New Orleans, USA

[6] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

[7] IBM Research, Yorktown Heights, USA

[8] Symbl.ai, Seattle, USA

[9] University of Washington, Seattle, USA