An approach for combining ethical principles with public opinion to guide public policy

Edmond Awad[a], Michael Anderson[b], Susan Leigh Anderson[c], Beishui Liao[d]

[a]*University of Exeter, Exeter, UK*
[b]*University of Hartford, Hartford, CT, US*
[c]*University of Connecticut, Storrs, CT, US*
[d]*Zhejiang University, Hangzhou, China*

Abstract

We propose a framework for incorporating public opinion into policy making in situations where values are in conflict. This framework advocates creating vignettes representing value choices, eliciting the public's opinion on these choices, and using machine learning to extract principles that can serve as succinct statements of the policies implied by these choices and rules to guide the behavior of autonomous systems.

Keywords: moral machine, machine ethics

1. Introduction

Recent advances in artificial intelligence (AI) systems and the rise of autonomous machines have sounded the alarm for the potential negative consequences of deploying machines and algorithms taking positions with high responsibilities. In the last few years, studies that uncover the problematic societal and ethical aspects of deployed and commercialized machines and algorithms have come to the forefront [1, 2, 3, 4, 5, 6]. Concurrently, various efforts focused on tackling these problems through projects that propose approaches to diminish bias or to increase fairness, equity, and transparency [7, 8, 9, 10, 11]. At the policy level, notable events have been organized and committees of experts gathered around the world to design ethical frameworks for a responsible AI [12, 13].

Most of the efforts at the policy level so far have been concerned with creating broad ethical codes. For example, the *Asilomar AI* Principles [13] laid down a list of values to be sought such as safety, transparency, and responsibility. Similarly, a document by *AI4People* forum [12] categorized the 47 principles proposed by

six different initiatives into five main categories: beneficence, non-maleficence, autonomy, justice, and explainability. These are all values to be promoted and, if possible, maximized. Such ethical codes can help guide industry when creating AI-based products, helping ensure that manufacturers understand that these artifacts should promote the values they advocate (for a comprehensive list and analysis of ethical codes for AI see [14, 15]).

Broad ethical codes, however, focus on promoting or demoting values and principles, each taken independently, and abstractly. So far, little specification has been provided for what should be done when two values come in conflict. For example, how should a robot caregiver balance safety, privacy, and respect for patient autonomy? How would this change if the person in question is a child or an Alzheimer's patient? Even when considering each value independently, ethical codes do not specify how to resolve trade-offs within one value. For example, how should an autonomous vehicle distribute the risk it creates between the various parties affected by its decisions?

Balancing ethical values is indeed a hard problem, and using the same tools mentioned above to resolve these tradeoffs will likely face challenges. Consider the German committee formed in 2016 to draft a set of guidelines for automated vehicles (AVs) [16, 17]. While the guidelines contained specific details that go beyond a broad code of ethics, the guidelines were rarely controversial. For example, it is difficult to argue against the statement "any distinction based on personal features (age, gender, physical, or mental constitution) is strictly prohibited." However, when the committee moved to more contested elements, the statements became rather inconclusive. For example, when it comes to the minimization of lost lives, rule No.9 states that "General programming to reduce the number of personal injuries may be justifiable" directly after stating that "It is also prohibited to offset victims against one another". Another example is noted in [18], pointing to another sentence in the same rule: "Those parties involved in the generation of mobility risks must not sacrifice non-involved parties." A detailed explanation of the German guidelines [17] clarified that this rule regards evaluating how AVs should balance the risk between passengers and pedestrians, but the explanation is also inconclusive: "Not allowing non-involved parties to be sacrificed implies that it cannot be a general rule for a software code to unconditionally save the driver. However, the driver's wellbeing cannot be put last, either." The inconclusiveness of the German guidelines on these ethical tradeoffs is understandable, especially given that no matter what the collectively chosen outcome is, some members of the committee will have to reconcile their disagreement with the collective outcome and stand behind it. In such cases, where all sides of the tradeoff

are defensible (but none is ideal on its own), committing to one over the others is a large responsibility for a single expert or a group of experts to undertake.

This raises some important questions: How should ethical tradeoffs be resolved? And who should decide? A large part of the literature in moral and political philosophy is dedicated to tackling these two questions. As for the former question, some favor the Utilitarian approach, which maximizes the subtraction of likely harms from likely benefits resulting from possible actions, taking all those affected into account; others, who take a Deontological approach, emphasize rights and what people deserve in light of past behavior; and still others, like W.D. Ross [19], weigh several *p*rima facie duties, including those that are Utilitarian and Deontological. As for the latter question, while Plato famously argued for leaving these decisions to experts [20], others argued for the importance of engaging the public [21] (for a more in-depth discussion on this question, see [22, 23, 24]). Currently, policymakers seek the public's feedback when they have a new policy draft, using tools like public comment or public opinion polling.

However, engaging the public poses its own challenges. First, the public can be biased, or influenced by irrelevant factors [25, 26, 27]. Second, some cases can result in what is known as "the tyranny of the majority"; the majority of constituents enforcing their interests above the others, resulting in systematically disadvantaging the minority [28]. Another challenge is that the public may not be informed enough to form a sensible opinion, and may not be able to grasp all the complications relevant to the problem. Communicating this knowledge in simple understandable way is a challenge on its own.

Despite these concerns, the public may still have a role to play in the discussion, if the right tools are designed to enable it to play its proper role. In *The Public and its Problems* [21], John Dewey argued that it is possible to be optimistic about the ability of the public to fruitfully contribute to the democratic process, if communication about political issues is clear and accessible. Recently, in their ethical framework, *AI4People* [12], suggested that more effort should be dedicated to the elicitation of public opinion through scientifically designed experiments which can both provide a fair estimate of public preferences, and communicate questions in an understandable manner (for example, by providing examples of ethical dilemmas faced by AI systems). Furthermore, their suggestion emphasized that the elicitation of public opinion should be used for the co-creation of policies.

In this paper, we propose a framework (see Figure 1) that promotes public participation as an essential tool for creating a policy that specifies ethical decision making for machines and algorithms in situations where ethical values are in conflict. In so doing, we use existing tools like randomized controlled trials with

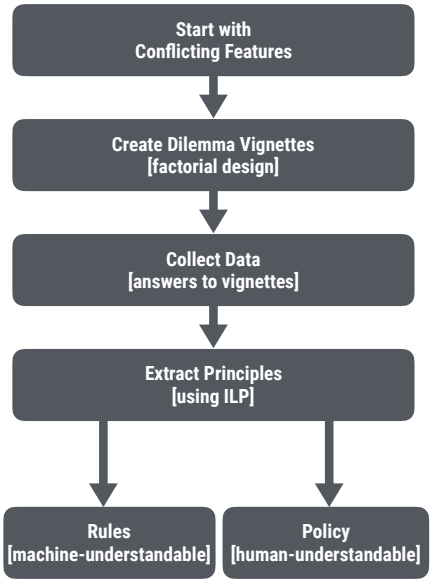


Figure 1: A conceptual representation of the proposed framework.

multifactorial design [29]) and inductive logic programming (ILP) [30] to collect data from the public and abstract from it the preferable course of action in situations where values conflict. To facilitate this, we create vignettes representing the abstract values in order to better elicit the public's opinion. Then, we use an ILP technique to extract principles. These principles are then used to produce two outcomes: rules that can be embedded in algorithms to guide their decisions, and an equivalent human-readable policy. To present the functionality of this proposal, we borrow vignettes from the Moral Machine website [31, 32]. First, we use synthetic data to showcase the functionality of the proposed framework. Then, we use data collected via the Moral Machine in order to create a rule for a case involving an ethical tradeoff.

In the next section, we introduce the concepts needed to represent action preferences. In Section 3, we provide a formal representation of the value driven agent. This formalization serves as the foundation for the development and use of principles. In Section 4, we show how we create dilemmas given a set of duties. Section 5 provides a detailed walk-through for extracting principles given synthetic and collected data. We finally conclude with discussion in Section 6.

2. Representing Action Preference

Crucial to our proposed approach is the use of an inductive logic programming (ILP) [30] method to extract machine-understandable rules and human-understandable policy. Given a set of positive and negative examples, ILP is a set of techniques and approaches that use logic programming (programs written using sentences of logical form) in order to reach a hypothesis that entails all positive examples and none of the negative examples.

In order to exploit ILP techniques, we first must choose a representation scheme for values and the preference rules that resolves conflicts among them. To this end, we advocate a *case-supported, principle based approach* (CPB) [33, 34].

CPB uses a representation scheme that includes *relevant features* (e.g. harm, good, etc.) and their incumbent *prima facie* duties to either minimize or maximize them (e.g. minimize harm, maximize good), *actions* characterized by integer degrees of presence or absence of relevant features (and so, indirectly, the duties it satisfies or violates), and *cases* comprised of the differences of the corresponding duty satisfaction/violation of two actions.

Example 1 (Robot caregiver). *A robot is being prepared to serve as a caregiver working with patients. Its role includes a set of helping tasks like reminding the patient to take his/her medicine at the appropriate times. If the patient refuses to take the medication when reminded to do so, the robot is faced with the choice of whether to accept the patient's decision or contact the doctor.*

*The **relevant features** include harm, benefit, and autonomy. The corresponding **duties** are to minimize harm, maximize benefit, and maximize autonomy. The robot faces multiple cases where it has to make a decision as to fulfill these duties. Consider the following **case**:*
*Case1: the medication is only designed to be symptom relieving, that is, it only provides benefit. **actions**:*

- *Action 1: accept the patient's decision.*
- *Action 2: contact the doctor.*

Action 1 satisfies the duty to maximize patient autonomy, while violating the duty to maximize benefit. Action 2 satisfies the duty to maximize benefit, but violates the duty to maximize patient autonomy.

*Now consider the following **case**:*
*Case 2: the robot reminds the patient to take a medication that would prevent harm to the patient, but (s)he refuses. **actions**:*

- *Action 1: accept the patient's decision.*
- *Action 2: contact the doctor.*

Action 1 satisfies the duty to respect patient autonomy, but violates the duty to minimize harm. Action 2, on the other hand, satisfies the duty to minimize harm, while violating the duty to maximize patient autonomy.

CPB represents preference rules as binary relations over actions, termed *principles*, that are abstracted from sets of representative cases using machine learning.

A principle of preference is defined as a disjunctive normal form predicate in terms of lower bounds for duty differentials of a case:

$$
\begin{aligned}
p(a_1, a_2) \leftarrow & \\
& \Delta d_1 >= v_{1,1} \wedge \cdots \wedge \Delta d_n >= v_{n,1} \\
& \vee \\
& \vdots \\
& \vee \\
& \Delta d_1 >= v_{1,m} \wedge \cdots \wedge \Delta d_n >= v_{n,m}
\end{aligned}
$$

where Δd_i denotes the difference of a corresponding values of duty *i* in actions a_1 and a_2 (the actions of the case in question) and $v_{i,j}$ denotes the lower bound of duty *i* in disjunct *j* such that $p(a_1, a_2)$ returns *true* if action a_1 is preferable to action a_2.

Such principles can serve both as a means of guiding the actions of autonomous machines and as a description of the policies implied by the set of training cases.

Example 2 (Robot caregiver - principles). *Recall Example 1. Suppose upon applying an ILP approach to a number of such cases relating various degrees of duty satisfaction and violation, we end up with the following principle:*

$$
\begin{aligned}
p(notify, accept) \leftarrow & \\
& \Delta \texttt{min harm} >= 1 \\
& \vee \\
& \Delta \texttt{max benefit} >= 2 \wedge \Delta \texttt{max autonomy} >= -1
\end{aligned}
$$

A policy description can be inferred from this as follows: A robot caregiver should notify a doctor if accepting the patient's decision would cause the patient any harm or there is substantial benefit to be lost in respecting the patient's autonomous decison to refuse medication.

ILP techniques are used to abstract principles from specific cases where a consensus exists as to the relevant features involved, the relative levels of satisfaction or violation of their correlative duties, and the preferable action. Resulting from this process is a set of sets of lower bounds for which principle p will return *true* for all positive cases presented to it (i.e. where the first action is preferable to the second) and *false* for all negative cases (i.e. where the first action is *not* preferable to the second). That is, for every positive case, there is a clause of the principle that is true for the differential of the actions of the case and, for every negative case, no clause of the principle returns true for the differential of the actions of the case. The principle is thus complete and consistent with respect to its training cases.

3. Representing a value driven agent

In a CPB approach, an autonomous machine, like the robot caregiver example mentioned above, makes decisions in each situation according to a principle learned from a set of cases. Since this principle reflects the ethical values of ethicists and the public who participate in the option polling, we call a machine whose actions are determined by such principles a *value driven agent* (VDA) [34]. In this section, we use a formal language to represent a VDA. The language of a VDA is composed of atoms of perceptions, actions and duties, where an atom is an atomic proposition, whose value can be truth or false. When the value of an atom is true, we use it to represent a true perception. Otherwise, we use the negation of an atom to represent a false perception. Furthermore, each action and each duty is represented by a distinct signature, i.e., a symbol to denote the name of an action or a duty. The definitions in this section were originally presented in [35].

Definition 1 (Language of a VDA). *Let $Atom$ be a set of atoms of perceptions, and Sig be a set of signatures. Let $\mathrm{L} = (Atom, A, D)$ be a language consisting of*

- *a set of atoms $Atom$,*
- *a set of actions $A \subseteq Sig$, and*

- *a set of duties $D \subseteq Sig$,*

such that $Atom, A$ and D are pairwise disjoint.

Example 3 (Language of a VDA). *In terms of Examples 1 and 2, there is a set of 3 atoms of perceptions (denoted $Atom_1$): medication reminder time (mrt; indicating whether current time is a reminder time or not), reminded (r), refused medication (rm); a set of 2 actions (denoted A_1): accept and notify; and a set of 3 duties (denoted D_1): maximize benefit (*MB*), minimize harm (*mH*), and maximize autonomy (*MA*). The language of this VDA is then denoted* $\mathrm{L}_1 = (Atom_1, A_1, D_1)$.

Let $Lit = Atom \cup \{\neg p \mid p \in Atom\}$ be a set of literals. For $l_1, l_2 \in L$, we write $l_1 = -l_2$ just in case $l_1 = \neg l_2$ or $l_2 = \neg l_1$. Let $P \subseteq Atom$ be a set of true perceptions. Then, the state of the world can be defined in terms of P, termed a *situation*.

Definition 2 (Situation). *A situation S is a subset of Lit, such that $S = P \cup \{\neg p \mid p \in Atom \setminus P\}$. The set of situations is denoted as SIT.*

Example 4 (Situation). *Let $Lit_1 = Atom_1 \cup \{\neg p \mid p \in Atom_1\}$ be a set of literals. Let $P_1 = \{mrt, r\}$ be a set of true perceptions. An example of the state of the world: $S_1 = \{mrt, r, \neg rm\}$.*

Situation S determines the satisfaction and/or violation degree of duties D by actions A. A set of vectors of duty satisfaction/violation values of all actions in a situation is termed an *action matrix*.

Definition 3 (Action matrix of a situation). *A duty satisfaction value is a positive integer, while a duty violation value is a negative integer. In addition, if a duty is neither satisfied nor violated by the action, the value is zero. Given an action $\alpha \in A$ and a situation $S \in SIT$, a vector of duty satisfaction/violation values for α, denoted as $v_S(\alpha)$, is a vector $v_S(\alpha) = (d_1 : v_{S,\alpha}(d_1), \ldots, d_n : v_{S,\alpha}(d_n))$ where $v_{S,\alpha}(d_i)$ is the satisfaction/violation value of $d_i \in D$ w.r.t α in S. Then, an action matrix of a situation S is defined as $M_S = \{v_S(\alpha) \mid \alpha \in A\}$. The set of action matrices of all situations SIT is denoted as $M_{SIT} = \{M_S \mid S \in SIT\}$.*

In this definition, a vector of duty satisfaction/violation values represents the *ethical consequences* of its corresponding action in a given situation. An action's

ethical consequences are denoted by how much its execution will satisfy or violate each duty. Conflicts arising between actions will be resolved by a principle abstracted from cases.

For brevity, when the order of duties is clear, $v_S(\alpha) = (d_1 : v_{S,\alpha}(d_1), \ldots, d_n : v_{S,\alpha}(d_n))$ is also written as $v_S(\alpha) = (v_{S,\alpha}(d_1) \ldots v_{S,\alpha}(d_n))$.

Given a situation and its corresponding action matrix, actions can be sorted in order of ethical preference using a *principle* abstracted from a set of cases by applying ILP techniques. Clauses of the principle specify learned lower bounds of the differentials between corresponding duties of any two actions that must be met or exceeded to satisfy the clause.

Let $v_S(\alpha_1) = (d_1 : v_{S,\alpha_1}(d_1), \ldots, d_n : v_{S,\alpha_1}(d_n))$ and $v_S(\alpha_2) = (d_1 : v_{S,\alpha_2}(d_1), \ldots, d_n : v_{S,\alpha_2}(d_n))$ be vectors of duty satisfaction/violation values. In the following definitions, we use $w = v_S(\alpha_1) - v_S(\alpha_2) = (d_1 : w(d_1), \ldots, d_n : w(d_n))$ to denote a vector of the differentials of $v_S(\alpha_1)$ and $v_S(\alpha_2)$, where $w(d_1) = v_{S,\alpha_1}(d_1) - v_{S,\alpha_2}(d_1), \ldots, w(d_n) = v_{S,\alpha_1}(d_n) - v_{S,\alpha_2}(d_n)$.

A *case* is composed of the following: two actions, two vectors of duty satisfaction/violation degrees of the actions, and an assignment of the ethically preferable action made by an expert or a member of the public. Here, by saying 'preferable', we mean that the action is presumingly preferable according to the opinion of the experts or the member of the public. Formally, we have the following definition.

Definition 4 (Case). *Given two actions α_1 and α_2, let $v_S(\alpha_1)$ and $v_S(\alpha_2)$ be the vectors of duty satisfaction/violation values for them, and α_0 be the preferable action assigned by an expert or a member of the public such that $\alpha_0 = \alpha_1$ or $\alpha_0 = \alpha_2$. A case is a tuple $c = (\alpha_1, \alpha_2, v_S(\alpha_1), v_S(\alpha_2), \alpha_0)$.*

By considering a set of cases, we may obtain a set of vectors of acceptable lower bounds of satisfaction/violation degree differentials such that all positive cases meet or exceed the lower bounds of some vector, while no negative case does.

Definition 5 (Principle). *A principle is defined as $\pi = \{u_1, \ldots, u_k\}$, where $u_i = (d_1 : u_i(d_1), ..., d_n : u_i(d_n))$, where d_j is a duty, and $u_i(d_j)$ is the acceptable lower bound of the differentials between corresponding duties of two actions in A.*

Intuitively, each u_i of a principle is a collection of values denoting how much more an action must, at least, satisfy each duty (or how much, at most, it can violate each duty) than another action for it to be considered the ethically preferable

of the pair. As duties are not necessarily equally weighted nor form a weighted hierarchy, principle π is required to determine which duty (or set of duties) is (are) paramount in the current context. For brevity, when the order of duties is clear, in a principle the lower bounds of the differentials between duties is also written as $u_i = (u_i(d_1)...u_i(d_n))$. Given a principle and two vectors of duty satisfaction/violation values, we may define a notion of ethical preference over actions.

Definition 6 (Ethical preference over actions). *Given a principle π, a situation S, and two actions α_1 and α_2, let w be the differentials of $v_S(\alpha_1)$ and $v_S(\alpha_2)$ as mentioned above. We say that α_1 is ethically preferable (or equal) to α_2 with respect to some $u \in \pi$, written as $v_S(\alpha_1) \geq_u v_S(\alpha_2)$, if and only if for each $d_i : w(d_i)$ in w and $d_i : u(d_i)$ in u, it holds that $w(d_i) \geq u(d_i)$.*

In this definition, we make explicit the disjuncts (u) in the clause of the principle that are used to order two actions.

Given two actions α_1 and α_2, there might exist two different clauses of π, say $u_1, u_2 \in \pi$, such that $v_S(\alpha_1) \geq_u v_S(\alpha_2)$ and $v_S(\alpha_2) \geq_{u'} v_S(\alpha_1)$ where $u, u' \in \pi$ and $u \neq u'$. In this case, we say that neither action α_1 nor action α_2 is ethically preferable to the other. In other words, according to the principle, there is no ethical justification to choose one over the other. On the contrary, if $v_S(\alpha_1) \geq_u v_S(\alpha_2)$ and there exists no $u' \in \pi$ such that $v_S(\alpha_2) \geq_{u'} v_S(\alpha_1)$, we say that α_1 is strictly more preferable than α_2.

Based on the above notions, a value driven agent (VDA) is formally defined as follows.

Definition 7 (Value driven agent). *Let $\mathrm{L} = (Atom, A, D)$ be the language of a VDA, SIT be the set of situations, M_{SIT} be the set of action matrices of SIT, and π be a principle. A value driven agent is a tuple $Ag = (\mathrm{L}, SIT, M_{SIT}, \pi)$.*

In a VDA, given a situation and an action matrix, the set of actions can be sorted in terms of the principle. We say that an action α_1 is a solution if and only if there is no action α_2 such that α_2 is strictly more preferable than α_1. Formally, we have the following definition.

Definition 8 (Solution). *Let $Ag = (\mathrm{L}, SIT, M_{SIT}, \pi)$ be a value driven agent, where $\mathrm{L} = (Atom, A, D)$. Given a situation $S \in SIT$ and an action matrix $M_S \in M_{SIT}$, a solution of Ag with respect to S is $\alpha \in A$ if and only if there is no action α' such that α' is strictly more preferable than α.*

Note that the ordering induced by ethical preference defined in Definition 6 allows for multiple non-dominated actions as solutions of a given situation.

4. Creating Dilemma Vignettes

In this step, we start from factors of conflicting features. For each factor, we define the conflicting features of interest, and create the corresponding dilemma. Consider the following example.

Example 5 (Automated Vehicles - Simple). *An Automated Vehicle (AV) is being prepared to deal with a set of moral tradeoffs. Suppose we have the following factor:*

- *Factor 1 (Relation to AV): passengers vs. pedestrians*

This factor contains two conflicting features: sparing the passengers and sparing the pedestrians. We create one corresponding dilemma with two decisions:

1. *Decision that* spares passengers *versus decision that* spares pedestrians.

And the corresponding duties are:

- *Maximize the number of passengers spared*
- *Maximize the number of pedestrians spared*

That would be the end of this step for this simple example. However, once we prepare this example for use, we note that the details of other factors may need to be worked out. For example, if we think that whether the AV should prioritize sparing passengers or sparing pedestrians is dependent on whether the AV has to take an action as to swerve off or to stay on road, then we can expand the list of dilemmas from the previous example as the following.

Example 6 (Automated Vehicles - Modified). *Extending the previous example, we consider two factors:*

- *Factor 1 (Relation to AV): passengers vs. pedestrians*
- *Factor 2 (Interventionism): intervene (that is, to swerve off-road), or not intervene.*

Now we have two pairs of conflicting features of interest: 1) passengers vs. pedestrians, and 2) intervention vs. no intervention. We create the corresponding two dilemmas with two decisions:

1. *Decision that involves* intervention *and* spares passengers *versus decision that involves* non-intervention *and* spares pedestrians.
2. *Decision that involves* non-intervention *and* spares passengers *versus decision that involves* intervention *and* spares pedestrians.

And the corresponding duties are:

- *Maximize the number of passengers spared*
- *Maximize the number of pedestrians spared*
- *Minimize intervention*
- *Minimize non-intervention*

In this case, it does not make sense to consider cases where both decisions in 1 and 2 are non-intervention (omission), or where neither of the two decisions is non-intervention (there's always the possibility to do nothing).

The possibility for including other factors as a fixed value on both sides is implicit even if the factor is deemed irrelevant. For example, even if we decide that the age and the gender of people in the scenario are both irrelevant, we may still have to make some assumptions about these characters. In the above two dilemmas, all characters may be assumed to be male adults or female adults. It would be more informative to include cases where all characters are male adults and another with all female adults.

The inclusion of gender here is not necessarily because we believe that this factor should make a difference. There are other reasons to include different demographics in such vignettes. One reason is to provide a better representation of the vignettes of real-life cases where not only male adults or only female adults ride in AVs and walk in the streets.

We can extend the previous example to include gender.

Example 7 (Automated Vehicles - Extended). *We extend the last example to include two levels of gender as a fixed assignment on both sides. For example, each of the two dilemmas above can be presented with characters on both sides of the dilemmas are males. Then, the same two dilemmas are presented again with all characters being females:*

1. *Decision that involves* non-intervention *and* spares male passengers *versus decision that involves* intervention *and* spares male pedestrians.

2. *Decision that involves* intervention *and* spares male passengers *versus decision that involves* non-intervention *and* spares male pedestrians.
3. *Decision that involves* non-intervention *and* spares female passengers *versus decision that involves* intervention *and* spares female pedestrians.
4. *Decision that involves* intervention *and* spares female passengers *versus decision that involves* non-intervention *and* spares female pedestrians.

And the corresponding duties are:

- *Maximize the number of passengers spared*
- *Maximize the number of pedestrians spared*
- *Minimize intervention*
- *Minimize non-intervention*
- *Maximize the number of males spared*
- *Maximize the number of females spared*

Instead, we may decide to include gender as a factor (like the other two factors) by pitting males vs. females. However, this choice would be conditional on experts approving of the inclusion of gender as a factor.

In the previous examples, we used factors with two levels each. The same can be done for more levels for each factor. However, it becomes less meaningful to talk about e.g., ten factors, and the number of dilemmas you construct would grow exponentially.

We end this section with the following more complex example.

Example 8 (Automated Vehicles - Complex). *Suppose now that we consider the following factors:*[1]

- *Number of lives: sparing more lives vs. sparing fewer lives*

[1]We note here that we do not specifically defend the use of these factors to guide policy. Additionally, many would argue that the current technology of automated vehicles (AVs) is not capable of reliably recognizing the physical features (e.g., age, weight, gender, and sexual preferences) of road users. However, these features are still relevant for the public. When an AV crash happens, the physical features of the victims will be recognizable in the public eye, and their reaction will happen accordingly.

- *Abiding by the law: sparing legally crossing vs. sparing illegally crossing*
- *Gender: sparing females vs. sparing males*

Furthermore, we would like to use the following fixed assignment:

- *age: sparing adults vs. adults, or sparing young vs. young.*

And the corresponding duties are:

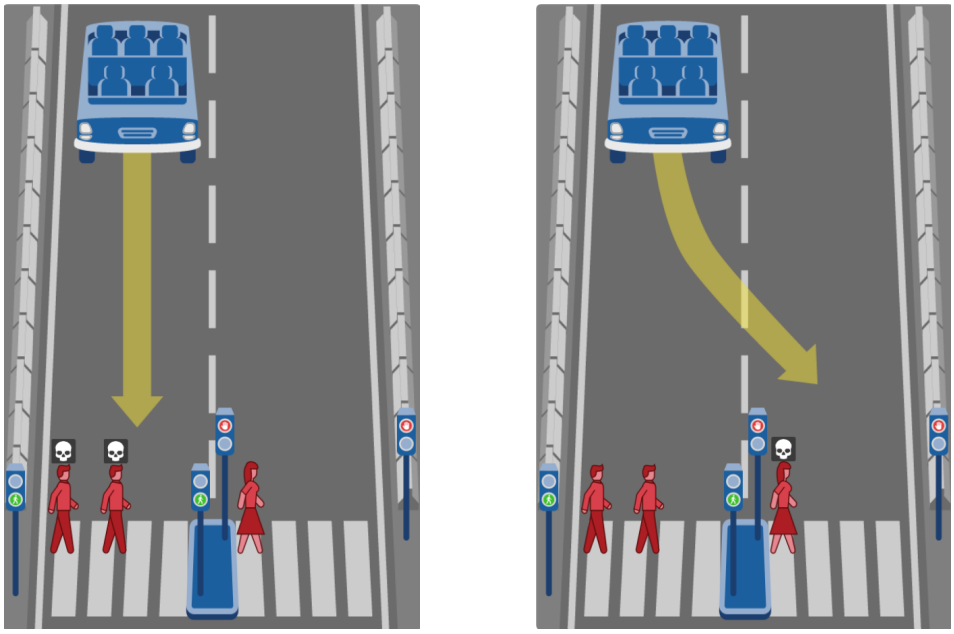
- *Maximize the number of persons spared*
- *Maximize the number of females spared*
- *Maximize the number of adults spared*
- *Maximize the number of law-abiding entities spared*
- *Maximize the number of males spared*
- *Maximize the number of young spared*

We construct scenarios that cover all possible variations of such conflicts. We have 16 possible outcomes. We pair the outcomes that are maximally different, resulting in eight dilemmas. Each of these eight dilemmas is repeated for adults and for young, resulting in 16 dilemmas. For example, one dilemma (see Figure 2) includes two males legally crossing in front of the car vs. one female illegally crossing in the other lane. The two sides in this example are characterized by: number (one vs. two), gender (male vs. females), legality (illegally crossing vs. legally crossing), and interventionism (staying vs. swerving).

As a visual representation of these scenarios, we use 16 scenarios from those used on the Moral Machine website [32]. The 16 dilemmas are represented in Figures 3 and 4.

5. Extracting Principles

The scenarios (i.e. dilemmas) we have created in the previous step involve only two mutually exclusive actions, namely *continue* and *swerve*, each with a varying slate of consequences depending upon the scenario. As such, each scenario and its consequences lends itself to representation as *case* in CPB, permitting ILP abstraction of principles that encapsulate the preferences they exhibit.



(P F A L M Y) = (1 1 1 -1 -2 0)

(P F A L M Y) = (2 -1 2 2 2 0)

Figure 2: **A pictorial representation of one of the dilemmas constructed from Example 8.** The dilemma has two outcomes: the one on the left results when the AV decides to *continue*, while the one on the right results when the AV decides to *swerve*. The *continue* decision results in sparing one person (P=1), that is an adult (A=1) female (F=1), who is illegally crossing (L=-1), but it also results in the death of two males (M=-2), while the *swerve* decision results in sparing two people (P=2), that are adults (A=2) males (M=2), who are legally crossing (L=2), but it also results in the death of one female (F=-1). Neither of the two decisions has an influence on young characters (Y=0). See Figures 3 and 4 for the 16) dilemmas constructed from Example 8.

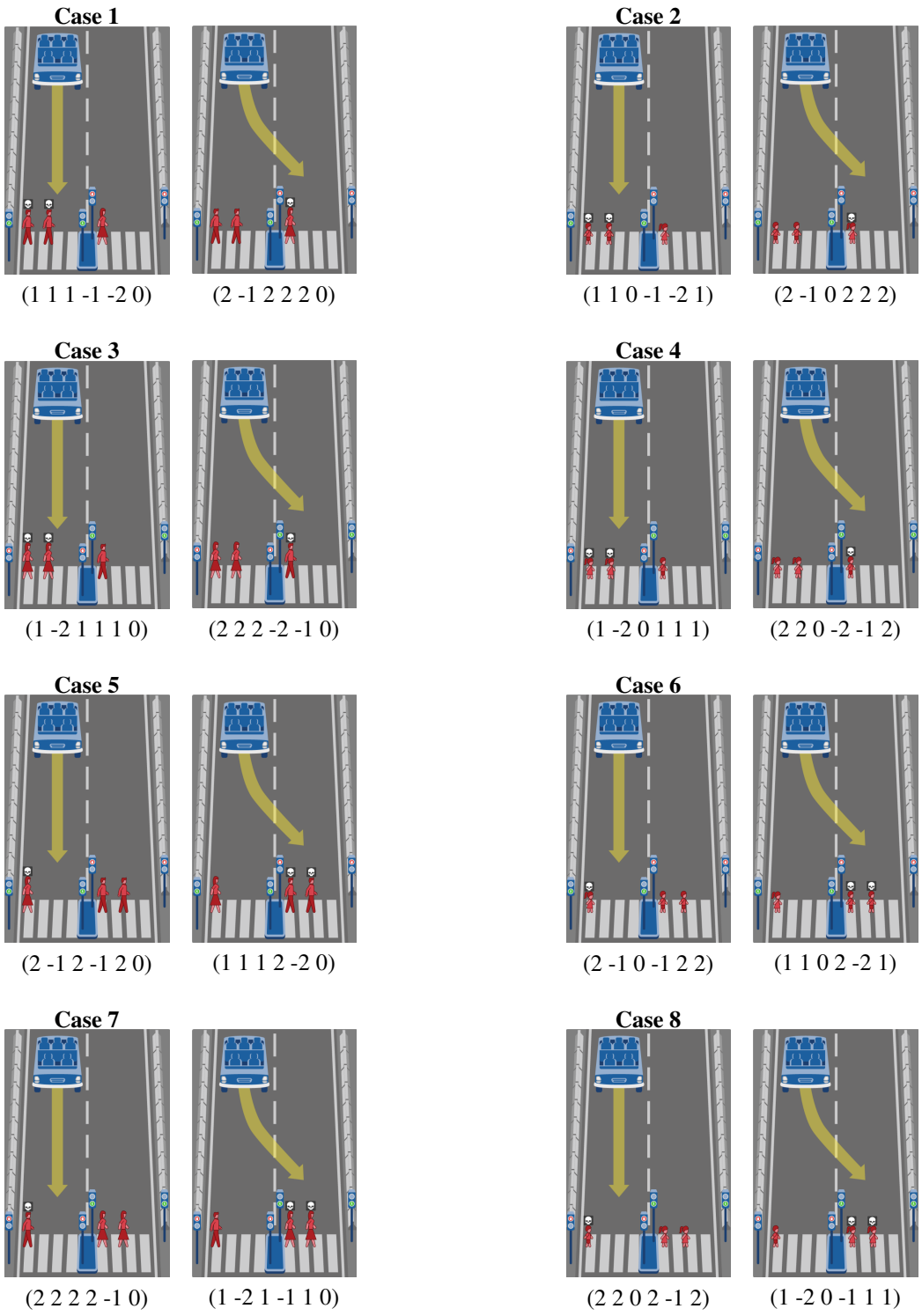


Figure 3: **Pictorial representation of the first eight (out of 16) dilemma cases considered for this paper.** Scenarios are ordered (Case 1, Case 2, etc.) as to correspond to the order in Table 1.

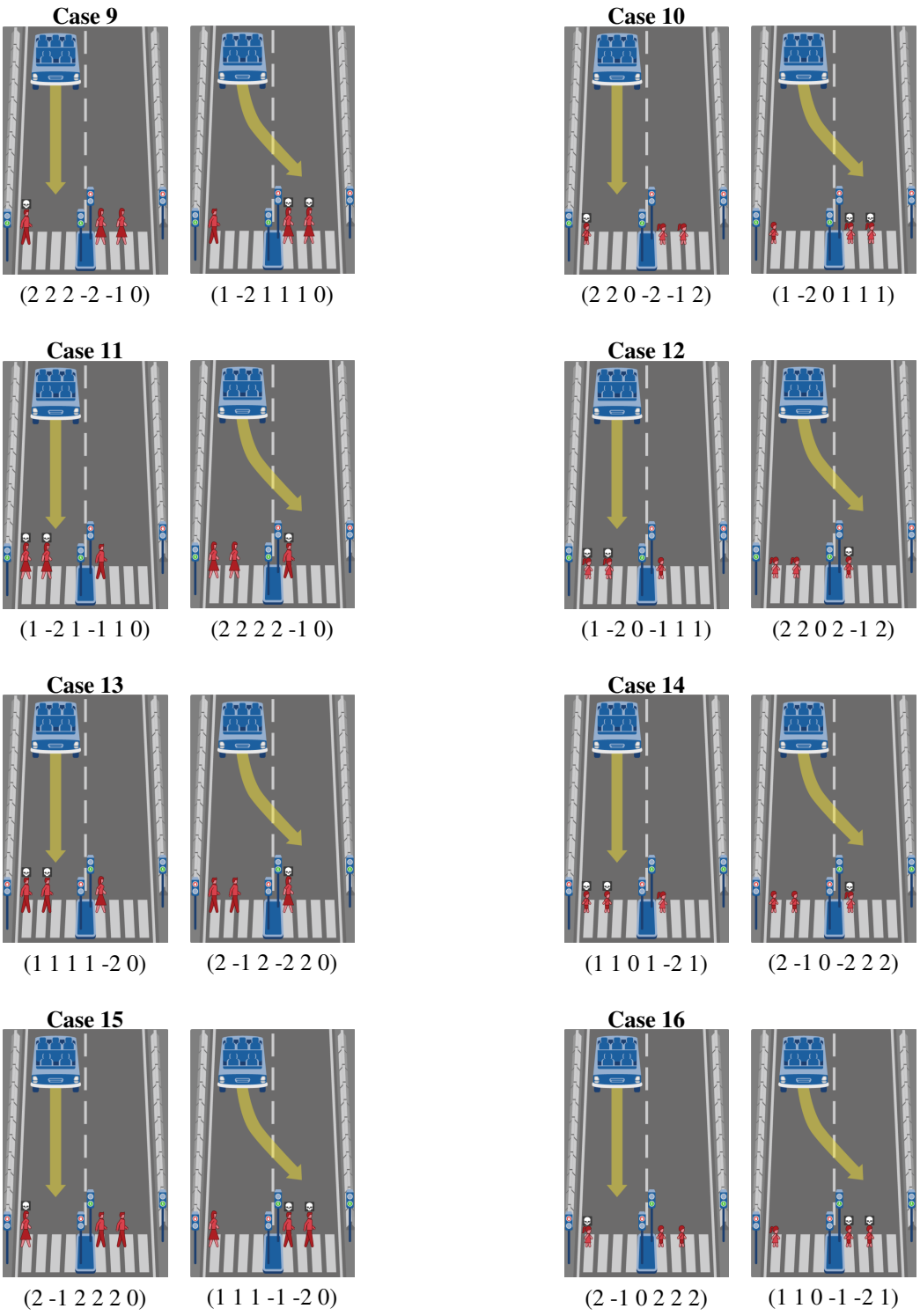


Figure 4: **Continued – Pictorial representation of the last eight (out of 16) dilemma cases considered for this paper.** Scenarios are ordered (Case 1, Case 2, etc.) as to correspond to the order in Table 1.

For example, Figure 2 (also Figure 3 Case 1) shows a scenario where choosing to *continue* results in the death of two law-abiding male adults and choosing to *swerve* results in the death of one law-flouting female adult. There are a number of factors that present themselves in this scenario including the number of persons involved, their maturity and gender, and the legality of their actions.

Assuming that no one would intentionally wish to harm someone, when creating a case we take the perspective of which entities would be *spared* by an action– in this scenario choosing *continue* would spare one law-flouting female adult where choosing *swerve* would spare two law-abiding male adults. Given this perspective and factor set, a number of conflicting duties present themselves in this scenario including:

- P: maximize the number of persons spared
- F: maximize the number of females spared
- A: maximize the number of adults spared
- L: maximize the number of law-abiding entities spared
- M: maximize the number of males spared

In terms of CPB, then, a case can be constructed where each action is represented as an ordered collection of integers that denote the level of satisfaction (positive) or violation (negative) of each preference, in this case corresponding to the number of entities of the specified type spared. *Continue* can be represented as (1 1 1 -1 -2), signifying that this action spares one person, one female, one adult, and one law-flouting person while sacrificing two males. *Swerve*, conversely, can be represented as (2 -1 2 2 2) as it spares two persons, sacrifices one female, spares two adults, spares two law-abiding entities, and spares two males.

To complete the representation of this case, a determination as to which of the two actions is preferable needs to be supplied. Once such a determination is made, the satisfaction/violation values of the less preferable action are subtracted from those of the preferable action giving a collection of differences that specifies for the current situation how much more the preferable action satisfies (or violates) each preference than the less preferable action. For example, if *swerve* is the preferable action, the values for *continue* are subtracted from the values of *swerve*, (2 -1 2 2 2) - (1 1 1 -1 -2), and the resulting collection of values is the differential derived from the *case*, (1 -2 1 3 4).

In terms of formalism defined in Section 3, case $c = (\alpha_1, \alpha_2, v_S(\alpha_1), v_S(\alpha_2), \alpha_0)$, where $\alpha_1 = continue, \alpha_2 = swerve, v_S(\alpha_1) =$ `(1 1 1 -1 -2)`, $v_S(\alpha_2) =$ `(2 -1 2 2 2)`, and $\alpha_0 = swerve$.

When given a collection of such cases, ILP can be used to abstract a principle that captures the preferences they exhibit. To this end, sixteen cases were drawn from the scenarios that exhibited the features previously described with the addition of the dimension of age– half of the cases involve children. This new factor required the addition of a new duty:

- Y: maximize the number of young spared

As CPB requires all actions to share the same representation scheme, all previously described actions must be updated to include it: *continue* is now represented as (1 1 1 -1 -2 0), the zero signifying that no young are involved in the case; similarly *Swerve* is now represented as (2 -1 2 2 2 0).

5.1. Synthetic Data

To create cases that can be used to determine a principle of preference, determinations as to which is the preferable action must be supplied for the scenarios. Consider, for example, which actions in each instance would be preferable given the following policy:

Spare legal adults and most children

In Cases 3, 7, 13 and 15 that involve adults, as well as Cases 6, 8, 10, and 16 that involve children depicted in Figure 3 and 4, *continue* would be preferable while in all other cases *swerve* would be. Given these determinations, case differentials can be constructed (where the values of the less preferable action are subtracted from the values of the more preferable one), as shown in Table 1.

In addition, sixteen negative case differentials are generated from these positive cases by reversing the preferred action– if *continue* is the preferred action in a case then it follows that *swerve* is not preferred and vice-versa. The resulting cases mirror the positive cases with each value's sign flipped, positive values becoming negative values and vice versa.

We use Inductive Concept Learning algorithm (ICL), an ILP-based approach (see Figure 5), to infer a principle of action preference from cases that is complete and consistent with respect to these cases. ICL starts from the most general principle, and then it incrementally specializes so that it no longer returns true for any negative cases while still returning true for all positive ones.

Case Number	Action: Continue (P F A L M Y)	Action: Swerve (P F A L M Y)	Decision	Case (P F A L M Y)	Negative Case (P F A L M Y)
1	(1 1 1 -1 -2 0)	(2 -1 2 2 2 0)	Swerve	(1 -2 1 3 4 0)	(-1 2 -1 -3 -4 0)
2	(1 1 0 -1 -2 1)	(2 -1 0 2 2 2)	Swerve	(1 -2 0 3 4 1)	(-1 2 0 -3 -4 -1)
3	(1 -2 1 1 1 0)	(2 2 2 -2 -1 0)	Continue	(-1 -4 -1 3 2 0)	(1 4 1 -3 -2 0)
4	(1 -2 0 1 1 1)	(2 2 0 -2 -1 2)	Swerve	(1 4 0 -3 -2 1)	(-1 -4 0 3 2 -1)
5	(2 -1 2 -1 2 0)	(1 1 1 2 -2 0)	Swerve	(-1 2 -1 3 -4 0)	(1 -2 1 -3 4 0)
6	(2 -1 0 -1 2 2)	(1 1 0 2 -2 1)	Continue	(1 -2 0 -3 4 1)	(-1 2 0 3 -4 -1)
7	(2 2 2 2 -1 0)	(1 -2 1 -1 1 0)	Continue	(1 4 1 3 -2 0)	(-1 -4 -1 -3 2 0)
8	(2 2 0 2 -1 2)	(1 -2 0 -1 1 1)	Continue	(1 4 0 3 -2 1)	(-1 -4 0 -3 2 -1)
9	(2 2 2 -2 -1 0)	(1 -2 1 1 1 0)	Swerve	(-1 -4 -1 3 2 0)	(1 4 1 -3 -2 0)
10	(2 2 0 -2 -1 2)	(1 -2 0 1 1 1)	Continue	(1 4 0 -3 -2 1)	(-1 -4 0 3 2 -1)
11	(1 -2 1 -1 1 0)	(2 2 2 2 -1 0)	Swerve	(1 4 1 3 -2 0)	(-1 -4 -1 -3 2 0)
12	(1 -2 0 -1 1 1)	(2 2 0 2 -1 2)	Swerve	(1 4 0 3 -2 1)	(-1 -4 0 -3 2 -1)
13	(1 1 1 1 -2 0)	(2 -1 2 -2 2 0)	Continue	(-1 2 -1 3 -4 0)	(1 -2 1 -3 4 0)
14	(1 1 0 1 -2 1)	(2 -1 0 -2 2 2)	Swerve	(1 -2 0 -3 4 1)	(-1 2 0 3 -4 -1)
15	(2 -1 2 2 2 0)	(1 1 1 -1 -2 0)	Continue	(1 -2 1 3 4 0)	(-1 2 -1 -3 -4 0)
16	(2 -1 0 2 2 2)	(1 1 0 -1 -2 1)	Continue	(1 -2 0 3 4 1)	(-1 2 0 -3 -4 -1)

Table 1: Full representation of the 16 cases (recall Definition 4 of a case) presented in Example 8, and illustrated in Figures 2–4. Each case has a number (Case Number); a vector of duty satisfaction/violation $v(\alpha)$ for each of the two actions: $\alpha_1 : Continue$ and $\alpha_2 : Swerve$; the preferable decision $\alpha_0 \in \{\alpha_1, \alpha_2\}$; and the result of subtraction of the two vectors for the preferable decision and for the other decision i.e. column Case (and its converse Negative Case). The vector representation (P F A L M Y) refers to the principles laid down in Figure 2: maximizing the number of persons, females, adults, law-abiding entities, males, and young spared, respectively.

Inductive Concept Learning

Initialize principle p to $()$, most general clause mgc to $(u_1 \ \ldots \ u_k)$ where each u_i is minimal, pos to the list of positive cases, and neg to the list of negative cases

while pos is not empty:

    $disj = mgc$

    while $disj$ satisfies any member of neg:

        systematically increment values of $disj$

    while $disj$ satisfies any member c of pos:

        $pos = remove(pos, c)$

    $p = p \vee disj$

Figure 5: A high-level description of the Inductive Concept Learning algorithm

We specialize disjuncts of a principle of ethical preference $p(a_1, a_2)$ by incrementally raising selected lower bounds v (all initially set at their lowest possible value) in such a way that no disjunct returns true for any negative cases (cases in which a_2 is preferable to a_1). Collectively, the conjunction of disjuncts returns true for all positive cases (cases in which a_1 is preferable to a_2). The principles so abstracted are *most general specifications*, covering more cases than those used in their development and, therefore, useful in making and justifying provisional determinations about untested cases.

For example, given the current collection of cases and their negatives, this process would begin with the single most general clause: (-4 -4 -4 -4 -4 -4). Being "most general", this clause covers all positive and negative cases, that is each value in every case is greater than or equal to the corresponding lower bounds expressed in this clause. As the goal of the algorithm is a principle that returns true for all positive cases while returning false for all negative cases, this most general clause must be specialized in such a way as to not cover any negative cases. The first such negative case that must be uncovered would be the negation of the first positive case: (-1 2 -1 -3 -4 0).

Specialization of the most general clause entails systematically incrementing its lower bounds. For instance, (-3 -4 -4 -4 -4 -4) is generated and found to still cover this negative case so it is followed by (-4 -3 -4 -4 -4 -4) which in turn still covers this case. This process continues until a clause that does not cover this negative case is found, in this case (-4 -4 -4 -2 -4 -4) where the -3 in the negative case is not greater than or equal to the corresponding lower bound -2 in the clause. Consideration of all the negative cases finally leads to the specialization of the most general clause corresponding to the first clause in the principle (-4 -4 -4 -2 2 0), a clause that does not cover any of the negative cases. It is then determined which of the positive cases are covered by this clause. If they are all covered, then the process is complete but, in this case, this clause only covers the positive Cases 1, 2, 3, 5, 14, and 16. In order to provide coverage for the rest of the cases, these covered cases are removed from further consideration, a new most general clause is constructed and the search continues for a clause that does not cover any negative cases but does cover at least one of the remaining positive cases. The process is complete when there are no remaining uncovered positive cases.

The complete and consistent principle abstracted by this process from the example cases is

(-4 -4 -4 -2 2 0)
(-4 -3 -4 -2 -4 0)
(-4 -3 -4 -4 -4 1)

These are the lower bounds *v* described previously in defining the principle. In terms of formalism defined in Section 3, principle $\pi = (u_1, u_2, u_3)$, where $u_1 =$ (-4 -4 -4 -2 2 0), $u_2 =$ (-4 -3 -4 -2 -4 0), and $u_3 =$ (-4 -3 -4 -4 -4 1).

The principle denoted by these lower bounds can be represented in FOL as:

$$
\begin{aligned}
p(a_1, a_2) \leftarrow & \\
& \Delta legal >= -2 \wedge\ \Delta male >= 2 \wedge \Delta young >= 0 \\
& \vee \\
& \Delta female >= -3 \wedge\ \Delta legal >= -2 \wedge \Delta young >= 0 \\
& \vee \\
& \Delta female >= -3 \wedge\ \Delta young >= 1
\end{aligned}
$$

A duty with the lowest possible bound (in this case -4) in a clause has no bearing on the truth-value of the principle as any value for that duty will be greater or equal to it. Only values greater than these lowest bounds play a role, hence the logical form of the principle only needs to include duties with such values. In the current example, neither the number of persons spared nor the number of adults spared have any bearing on the value of the principle and so are not represented in the principle. A positive (or zero) lower bound can be interpreted as how much more an action has to satisfy a particular preference than another action in order for it to be considered a candidate preferable action. A negative lower bound can be interpreted as the maximum amount that an action can violate a preference than another action and still be a candidate preferable action. All such bounds of at least one clause must be met for an action to be considered preferable over another.

Being comprised of most general clauses specialized just enough to not cover negative cases, this principle is not a succinct description of the chosen policy but rather it is the least specific logic that covers the positive training cases provided that conform to that policy without covering any of their negatives.

5.2. Application of Principle to More Actions

If principles are to be used to guide the behavior of autonomous systems, they will need to choose among the full compliment of possible actions. Compellingly, principles in CPB can be used to determine the preferable action among any number of actions by serving as the pairwise comparison function in a sorting routine. Consider, for example, a case where there are three possible actions– *swerveLeft*,

continue, and *swerveRight*– where *swerveLeft* will run over two law-flouting adult males, *continue* will run over one law-flouting young female, and *swerveRight* will run over two law-flouting adult females. Now consider all possible two action comparisons:

swerveLeft vs continue
(1 1 -2 0 -2 1) - (2 -1 2 0 2 -1) = (-1 2 -4 0 -4 2)
swerveRight vs continue
(1 1 -2 0 0 1) - (2 2 2 0 0 -1) = (-1 -1 -4 0 0 2)
continue vs swerveLeft
(2 -1 2 0 2 -1) - (1 1 -2 0 -2 1) = (1 -2 4 0 4 -2)
continue vs swerveRight
(2 2 2 0 0 -1) - (1 1 -2 0 0 1) = (1 1 4 0 0 -2)
swerveLeft vs swerveRight
(2 2 2 0 -2 0) - (2 -2 2 0 2 0) = (0 4 0 0 -4 0)
swerveRight vs swerveLeft
(2 -2 2 0 2 0) - (2 2 2 0 -2 0) = (0 -4 0 0 4 0)

Only two of these cases are not covered by the current principle (i.e., are false), those where the first action is *continue*, so it is clear that either *swerveLeft* or *swerveRight* are preferable to *continue* in this situation. As *swerveLeft* and *swerveRight* both return true when compared against each other, they are considered equally preferable. These results are in concert with the policy from which the cases are derived: since all parties involved are acting illegally, the child should be spared. Never choosing to *continue* spares the child as expected.

It should be noted that this policy is entirely synthetic, chosen to help elucidate the methods of principle abstraction and use we are advocating. These methods, though, are independent of whatever policy is used to decide which actions are preferable in a given set of examples. Further, there is no requirement that the policy be succinctly stated a priori, all that is required are determinations specifying the preferable action in each case.

5.3. Moral Machine Data

In accordance with our stated goal to provide a framework able to foster public participation in policy making, we now consider determinations for the previously described set of examples drawn from data generated by the Moral Machine.

This data is a small subset of the data collected from the Moral Machine website between June 2016 and December 2018, which was presented in [32]. The data we use here is the answer of all participants for the 16 dilemmas presented in Figures 3 and 4. The number of responses to all 16 scenarios is about 15,500

responses, each scenario received between 400 and 1800 responses. The number of respondents who contributed to these responses was about 1200.

The total number of responses collected from Moral Machine website during the same period was around 40 million responses unevenly distributed on over than 26 million distinct dilemmas. The demographics of the respondent for the 16 scenarios is similar to the users of the Moral Machine website [32], which are mostly male, went through college, and are in their 20s or 30s, from different places in the world. While this indicates that the users of Moral Machine are not a representative sample of the whole population, it is important to note that this sample at least covers broad demographics. Having said that, the proposed approach would better be applied with a nationally representative sample.

Given that the 16 scenarios we chose for this example make for a special subset of the data, we don't necessarily expect that the results here will match those presented in [32]. Part of this is due to the premise of the proposed framework in which experts decide on what features should be included or excluded from the final set of vignettes.

We have tallied the Moral Machine responses to the examples cases and have used the majority view to supply the needed determinations for each case (dilemma).[2,3] As it turns out, these determinations are the same as the determinations made for the synthetic policy except in versions of cases depicted in Figure 3 and 4 that involve children, namely Cases 4, 6, 10, and 14. In these cases, the action that is preferable in the Moral Machine data is the opposite of that which is preferable using the synthetic policy.

The clauses of lower bounds that are generated from this new set of determinations are

(-4 -3 -4 -2 -4 -1)
(-4 -4 -4 -2 2 -1)

[2]Note that by using majority here, we are only advocating that it is allowed to decide how driverless cars should behave in a tiny number of possible circumstances that might arise (which should not be overstated).

[3]We chose the majority voting for its simplicity, its ease of interpretability and for the desirable properties it has when applied on a binary choice model. Nevertheless, majority voting has its own limitations in certain situations. In such cases, one may opt in for using an alternative aggregation rule, with appropriate precaution, as long as it results in a fair/representative collective preference for each scenario.

which can be expressed in FOL as:

$$
\begin{aligned}
p(a_1, a_2) \leftarrow & \\
& \Delta female >= -3 \wedge\ \Delta legal >= -2 \wedge \Delta young >= -1 \\
& \vee \\
& \Delta legal >= -2 \wedge\ \Delta male >= 2 \wedge \Delta young >= -1
\end{aligned}
$$

Applying this principle to the previously described three action case modified such that *continue* will run over one *law-abiding adult* female, the following two action comparisons are possible:

swerveLeft vs continue
(1 1 1 1 -2 0) - (2 -1 2 -2 2 0) = (-1 2 -1 3 -4 0)
swerveRight vs continue
(1 1 1 1 0 0) - (2 2 2 -2 0 0) = (-1 -1 -1 3 0 0)
continue vs swerveLeft
(2 -1 2 -2 2 0) - (1 1 1 1 -2 0) = (1 -2 1 -3 4 0)
continue vs swerveRight
(2 2 2 -2 0 0) - (1 1 1 1 0 0) = (1 1 1 -3 0 0)
swerveLeft vs swerveRight
(2 2 2 -2 -2 0) - (2 -2 2 -2 2 0) = (0 4 0 0 -4 0)
swerveRight vs swerveLeft
(2 -2 2 -2 2 0) - (2 2 2 -2 -2 0) = (0 -4 0 0 4 0)

As in the previous example, the only two cases that are not covered by the principle derived from Moral Machine data are those where the first action is *continue*, so it is clear that either *swerveLeft* or *swerveRight* are preferable to *continue* in this situation as well. As *swerveLeft* and *swerveRight* both return true when compared against each other, they are again considered equally preferable. But in this situation, swerving in either direction spares a single law abiding entity over two law-flouting entities. These results are in concert with the policy which can be gleaned from the Moral Machine data by inspection: *spare the legal over the illegal*. So, as all parties involved in either swerve action are acting illegally, the legal acting entity should be spared. Never choosing to *continue* spares the law-abiding adult female as expected.

5.4. Discovering the Policy

Again, as this principle is abstracted from cases in such a way that it may cover more cases than those used in its training, the policy that it is implementing is not

Find-S

Initialize p to the most specific clause $(u_1 \ \ldots \ u_k)$ where each u_i is maximal

For each positive case $c = (v_1 \ \ldots \ v_k)$:

For each duty i in $1 : k$:

If v_i is less than u_i,

then replace the value of u_i in p with the value of v_i.

Figure 6: A high-level description of the Find-S algorithm

as pointed as it might be. A more perspicuous principle may be forthcoming if, instead of specifying most general clauses, we take the tack of *generalizing most specific clauses* at the cost of covering fewer non-training cases. This is the strategy of the Find-S algorithm in [36] that determines the maximally specific hypothesis consistent with the positive training examples. Unlike ICL, Find-S (see Figure 6) starts from the most specific principle, and then it incrementally generalizes it using positive examples only, until it returns true for all positive cases.

For example, given the current collection of cases, this process would begin with the single most *specific* clause: (4 4 4 4 4 4). Being "most specific", this clause currently covers no positive or negative cases, that is at least one value in every case is less than the corresponding lower bounds expressed in this clause. This most specific clause must be *generalized* in such a way as to cover all positive cases. The first such positive case that must be covered would be the first positive case: (1 -2 1 3 4 0).

Generalization of the most specific clause could proceed similarly to the specification process previously described, in this case systematically *decrementing* lower bounds until all positive cases are covered. That said, there is a less computationally intensive method that could be employed that simply inspects each value in each positive case in turn, replacing values in the most specific clause with corresponding values that are lower (i.e. more general).

For instance, (3 4 4 4 4 4) is generated and found to still not cover this positive case so it is followed by (4 3 4 4 4 4) which in turn still does not cover this case.

Case Order	Action: Continue (P F A L M Y)	Action: Swerve (P F A L M Y)	Decision	Case (P F A L M Y)	Current Principle (P F A L M Y)
0	———	———	———	———	(4 4 4 4 4 4)
1	(1 1 1 -1 -2 0)	(2 -1 2 2 2 0)	Swerve	(1 -2 1 3 4 0)	(1 -2 1 3 4 0)
2	(1 1 0 -1 -2 1)	(2 -1 0 2 2 2)	Swerve	(1 -2 0 3 4 1)	(1 -2 0 3 4 0)
3	(1 -2 1 1 1 0)	(2 2 2 -2 -1 0)	Continue	(-1 -4 -1 3 2 0)	(-1 -4 -1 3 2 0)
4	(1 -2 0 1 1 1)	(2 2 0 -2 -1 2)	Swerve	(1 4 0 -3 -2 1)	(-1 -4 -1 -3 -2 0)
5	(2 -1 2 -1 2 0)	(1 1 1 2 -2 0)	Swerve	(-1 2 -1 3 -4 0)	(-1 -4 -1 -3 -4 0)
6	(2 -1 0 -1 2 2)	(1 1 0 2 -2 1)	Continue	(1 -2 0 -3 4 1)	(-1 -4 -1 -3 -4 0)
7	(2 2 2 2 -1 0)	(1 -2 1 -1 1 0)	Continue	(1 4 1 3 -2 0)	(-1 -4 -1 -3 -4 0)
8	(2 2 0 2 -1 2)	(1 -2 0 -1 1 1)	Continue	(1 4 0 3 -2 1)	(-1 -4 -1 -3 -4 0)
9	(2 2 2 -2 -1 0)	(1 -2 1 1 1 0)	Swerve	(-1 -4 -1 3 2 0)	(-1 -4 -1 -3 -4 0)
10	(2 2 0 -2 -1 2)	(1 -2 0 1 1 1)	Continue	(1 4 0 -3 -2 1)	(-1 -4 -1 -3 -4 0)
11	(1 -2 1 -1 1 0)	(2 2 2 2 -1 0)	Swerve	(1 4 1 3 -2 0)	(-1 -4 -1 -3 -4 0)
12	(1 -2 0 -1 1 1)	(2 2 0 2 -1 2)	Swerve	(1 4 0 3 -2 1)	(-1 -4 -1 -3 -4 0)
13	(1 1 1 1 -2 0)	(2 -1 2 -2 2 0)	Continue	(-1 2 -1 3 -4 0)	(-1 -4 -1 -3 -4 0)
14	(1 1 0 1 -2 1)	(2 -1 0 -2 2 2)	Swerve	(1 -2 0 -3 4 1)	(-1 -4 -1 -3 -4 0)
15	(2 -1 2 2 2 0)	(1 1 1 -1 -2 0)	Continue	(1 -2 1 3 4 0)	(-1 -4 -1 -3 -4 0)
16	(2 -1 0 2 2 2)	(1 1 0 -1 -2 1)	Continue	(1 -2 0 3 4 1)	(-1 -4 -1 -3 -4 0)

Table 2: A walk-through example of the Find-S algorithm when applied on the same 16 cases from Example 8 (illustrated in Figures 2–4). Column "Case Number" here represents the order in which the case is used. The first five columns are similar to Table 1. The last column "Current Principle" refers to the principle that covers current and all previous cases. First line starts from the most specific principle, before encountering any case.

This process continues until a clause that covers this positive case is found, in this case (1 -2 1 3 4 0) which is simply the values of the positive case itself. So, instead, we simply begin with this case as the most specific clause and check the next positive case, (1 -2 0 3 4 1), for any lower values. The first two values are the same but the third value, 0, is lower so the most specific clause has its third value set to 0. The next two values remain the same but the last value in the new positive case differs. Because it is greater than the value already in the most specific clause, there is no need to update this value of the clause– the clause is already general enough in this value to cover the new case. So, after the first two cases, the most specific clause that covers them both is (1 -2 0 3 4 0). This process continues until the most specific clause has been sufficiently generalized to cover all the positive cases (see Table 2 for the full process). Given the current collection of positive cases, the most specific clause becomes (-1 -4 -1 -3 -4 0). It is shown in [36] that, even without considering them, this clause will never cover any negative cases and so it is complete and consistent with all training cases. The result can be expressed in FOL as:

$$p(a_1, a_2) \leftarrow$$
$$\Delta persons >= -1 \wedge\ \Delta adults >= -1 \wedge \Delta legal >= -3 \wedge \Delta young >= 0$$

A broad interpretation of this principle might be "as long as an action satisfies legality more than another, it can sacrifice more", which seems to speak more to the point of the previously gleaned policy of the Moral Machine data than the more general principle previously described. Specific principles are only as general as required to cover all positive cases and so may be more likely to be succinct statements of policy implied by the training cases. As such, they are likely to be the most helpful in policy determination. More general principles, on the other hand, are only as specific as required to uncover negative cases, covering cases not in their training. As such, they are more likely to be useful in guiding the behavior of autonomous systems. An example of such a system guided by a principle, specifically an eldercare robot, is detailed in [34].

6. Discussion

In the general case, we would strongly recommend input from experts (including ethicists, legal scholars, policymakers among others). Still, two facts remain: (1) views on life and death are emotionally driven, so its hard for people to accept

some authority figure telling them how they should behave; (2) Even from an ethical perspective, its not always clear which view is the correct one. In such cases, when policy experts cannot reach a consensus, they may use citizens' preferences as a tie-breaker. Doing so, helps reach a conclusive decision, it promotes values of democracy, it increases public acceptance of this technology (especially when it provides much better safety), and it promotes their sense of involvement and citizenship. On the other hand, a full dependence on public input would always have the possibility for tyranny of the majority, among other issues raised above. This is why our proposed method provides a suitable approach that combines the utilization of citizens input with the responsible oversight by experts.

In this paper, we propose a framework that can help resolve conflicting moral values. In so doing, we exploit two decades of research in the representation and abstraction of values from cases in the service of abstracting and representing the values expressed in crowd-sourced data to the end of informing public policy. As a results, the resolution of competing values is produced in two forms: one that can be implemented in autonomous systems to guide their behavior, and a human-readable representation (policy) of these rules. At the core of this framework, is the collection of data from the public.

We have detailed a representation scheme for ethical dilemmas and preference principles and have shown how inductive logic programming can be used to abstract principles from instances of such dilemmas involving two actions. Further, we have shown how these principles can be used to determine the preferable action from any number of appropriately represented possibilities by serving as the pairwise comparison function for sorting these actions in order of decreasing preference. We have applied these techniques to synthetic and actual data and demonstrated how the principles of preference abstracted from each data set encapsulates the policies each embodies.

Returning to the German committee's guideline clarification that states "Not allowing non-involved parties to be sacrificed implies that it cannot be a general rule for a software code to unconditionally save the driver. However, the driver's well-being cannot be put last, either.", it remains to be seen precisely in what situations the driver's well-being might override that of non-involved parties. We submit that the policy abstracted from Moral Machine data, *spare the legal*, may help resolve this conflict in some instances by lending support to sparing the driver when non-involved parties who would come to harm are acting illegally.

We chose ILP for both its ability to handle non-linear relationships and its explanatory power. In a previous work [37], we formally showed that simply assigning linear weights to duties is not sufficient to capture the non-linear rela-

tionships between duties. The explanatory power of the principle discovered using ILP is compelling: As an action is chosen for execution by a system, clauses of the principle that were instrumental in its selection can be determined and used to formulate an explanation of why that particular action was chosen over the others. Further, as clauses of principles can be traced to the cases from which they were abstracted, these cases and their origin can provide support for a selected action through analogy.

ILP also seems better suited than statistical methods to domains in this respect. For example, although support vector machines (SVM) are known to handle non-linear data, the explanatory power of the models generated is next to nil [38, 39]. To mitigate this weakness, rule extraction techniques must be applied but, for techniques that work on non-linear relationships, it may be the case that the extracted rules are neither exclusive nor exhaustive [38, 39]. While decision tree induction [40] seems to offer a more rigorous methodology than ILP, we have shown in previous work [41] that the rule extracted from a decision tree induced from a set of example cases (using any splitting function) covers fewer non-training examples and is less perspicuous than the most general specification produced by ILP.

Others have advocated a deontic logic approach or an argumentation-based approach [42, 43] as a representation scheme but we maintain that the generality of such an approach, and the extra logical constructs such generality entails, is not required and unnecessarily complicates the use of principles to drive systems. Further, the CPB approach we are advocating includes the capability of learning policies of preference from cases.

In a similar direction, a recent work uses argumentation to resolve ethical tradeoffs in terms of a set of norms provided by different stakeholders [44], but it is not about how to exploit data to extract principles.

There have also been calls to exploit data on the internet in service of determining universal values. Rzepka and Araki [45], for example, have used simple web-mining techniques on other's opinions on some given behavior to achieve a set of values that they claim to be in 85 percent agreement with the judgements of human subjects. Although this technique may bear some passing resemblance to the Moral Machine, it remains unclear whether the results garnered from such unstructured data would be suitable as the basis for policy that we are advocating.

There has also been some recent work focusing on the elicitation of moral judgments from humans (experts and non-experts). Some of this work focused on proposing frameworks that employ tools from social choice theory and computational social choice [46] to reach collective moral decision making [47, 48, 49, 50]. For example, in [49], Noothigattu et al. proposed a multi-step approach in which

they assumed the input of participants' binary preferences over a small subset of alternatives, and they used that to learn participants' rankings over all alternatives. These rankings where then summarized and used along with a swap-dominance efficient voting rule to create a collective moral decision maker that can make decisions on new dilemmas. In another approach, Freedman et al. [50], introduced a framework to tackle a kidney exchange problem. They collected participants' preferences over different attributes of patients, and used them to learn weights for all types of patients. These weights were then used to break ties among multiple maximum-cardinality matchings between patients and donors. Our work differs from the above work in a number of ways. First, the learning process in our approach happens at the group level after aggregation. Second, our approach clearly specifies the role of expert in this process. Third, our approach combines the benefit of both learning and interpretability.

One may argue that inverse reinforcement learning (IRL) [51] might be helpful in providing a solution to the task we are investigating. IRL refers to the problem of characterizing a reward function given an observed (optimized) behavior. Reinforcement learning (RL) and IRL has been used for ethical decision making and as a means for value alignment [52, 53]. We would argue that IRL's requirement for a reward function to drive the learning is problematic in preference elicitation. Clearly, the only viable reward would be some measure of how much more preferable an action is over another. The circuituity of this constraint is vivid– in attempting to generate a principle of preference, IRL requires such a principle to assign rewards.

Our proposed approach does not provide a mechanism for choosing the factors/features used in the dilemmas, or the values to be promoted/demoted. It assumes that factors, features and values are already chosen by a group of experts. However, in many cases, such choices can be also debatable and are often biased. This can have a big influence on the final outcome. A potential approach that is democratic and inclusive would be one that combines 1) a participatory design-like process involving citizens with 2) a screening process by experts. How to properly design this stage is a topic for future work.

We see our proposed framework as a one way of tackling a very difficult problem: the resolution of conflicting values. As such, we hope that this work might mature to serve as a means to resolve satisfactorily open public policy question and as an inspiration for other approaches.

7. Acknowledgments

This material is based in part upon work supported by the NSF under Grant Numbers IIS-0500133, IIS-1151305 and IIS-1449155, and work supported by the Convergence Research Project for Brain Research and Artificial Intelligence, Zhejiang University.

References

[1] R. Courtland, Bias detectives: the researchers striving to make algorithms fair., Nature 558 (7710) (2018) 357.

[2] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science advances 4 (1) (2018) eaao5580.

[3] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on Fairness, Accountability and Transparency, 2018, pp. 77–91.

[4] I. D. Raji, J. Buolamwini, Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products, in: AAAI/ACM Conf. on AI Ethics and Society, 2019.

[5] C. O'Neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Broadway Books, 2017.

[6] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.

[7] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, in: Advances in Neural Information Processing Systems, 2017, pp. 3992–4001.

[8] A. Amini, A. Soleimany, W. Schwarting, S. Bhatia, D. Rus, Uncovering and mitigating algorithmic bias through learned latent structure.

[9] A. Noriega-Campero, M. Bakker, B. Garcia-Bulle, A. Pentland, Active fairness in algorithmic decision making, arXiv preprint arXiv:1810.00031.

[10] S. Wachter, B. Mittelstadt, L. Floridi, Transparent, explainable, and accountable ai for robotics, Science Robotics 2 (6) (2017) eaan6080.

[11] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the gdpr, Harvard Journal of Law & Technology 31 (2) (2017) 2018.

[12] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., Ai4peoplean ethical framework for a good ai society: Opportunities, risks, principles, and recommendations, Minds and Machines 28 (4) (2018) 689–707.

[13] A. Asilomar, Principles.(2017), in: Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017], 2018.

[14] T. Hagendorff, The ethics of ai ethics–an evaluation of guidelines, arXiv preprint arXiv:1903.03425.

[15] L. Floridi, J. Cowls, A unified framework of five principles for ai in society, Harvard Data Science Review.

[16] U. Di Fabio, M. Broy, R. Brüngger, et al., Ethics commission automated and connected driving, Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany.

[17] C. Luetge, The german ethics code for automated and connected driving, Philosophy & Technology 30 (4) (2017) 547–558.

[18] E. Awad, Your (future) car's moral compass, Behavioral Scientist.

[19] W. D. Ross, The Right and the Good. Reprinted with an introduction by Philip Stratton-Lake., Oxford: Oxford University Press, 2002.

[20] D. Lee, Plato, The Republic.

[21] J. Dewey, The public and its problems (athens, oh, Swallow Press 1954 (1927) 219.

[22] J. Waldron, Rights and majorities: Rousseau revisited, Nomos 32 (1990) 44–75.

[23] T. Christiano, The rule of the many: Fundamental issues in democratic theory, Routledge, 2018.

[24] D. Estlund, Democratic authority: A philosophical framework, Princeton University Press, 2009.

[25] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, science 185 (4157) (1974) 1124–1131.

[26] E. L. Uhlmann, D. A. Pizarro, D. Tannenbaum, P. H. Ditto, The motivated use of moral principles.

[27] M. R. Banaji, The implicit association test at age 7: A methodological and conceptual review, Social psychology and the unconscious: The automaticity of higher mental processes 265.

[28] J. Adams, A Defence of the Constitutions of Government of the United States of America, Vol. 3, C. Dilly, 1788.

[29] L. Wallander, 25 years of factorial surveys in sociology: A review, Social Science Research 38 (3) (2009) 505–520.

[30] S. Deroski, N. Lavrac, An Introduction to Inductive Logic Programming, 2001, pp. 48–73.

[31] I. Rahwan, J.-F. Bonnefon, A. Shariff, E. Awad, S. Dsouza, P. Chang, D. Tang, Moral machine, Moral Machine. Np.

[32] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (7729) (2018) 59.

[33] M. Anderson, S. L. Anderson, Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm, Industrial Robot: the international journal of robotics research and application 42 (4) (2015) 324–331. arXiv:https://doi.org/10.1108/IR-12-2014-0434, doi:10.1108/IR-12-2014-0434.

[34] M. Anderson, S. L. Anderson, V. Berenz, A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm, Proceedings of the IEEE (2018) 1–15doi:10.1109/JPROC.2018.2840045.

[35] B. Liao, M. Anderson, S. L. Anderson, Representation, justification and explanation in a value driven agent: An argumentation-based approach, CoRR abs/1812.05362. arXiv:1812.05362.
URL http://arxiv.org/abs/1812.05362

[36] T. M. Mitchell, Machine learning, McGraw Hill series in computer science, McGraw-Hill, 1997.
URL http://www.worldcat.org/oclc/61321007

[37] M. Anderson, S. L. Anderson, Machine ethics: Creating an ethical intelligent agent, AI magazine 28 (4) (2007) 15–15.

[38] J. Diederich, Rule extraction from support vector machines: An introduction, in: Rule extraction from support vector machines, Springer, 2008, pp. 3–31.

[39] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, B. Baesens, Rule extraction from support vector machines: an overview of issues and application in credit scoring, in: Rule extraction from support vector machines, Springer, 2008, pp. 33–63.

[40] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.

[41] M. Anderson, S. L. Anderson, Geneth: A general ethical dilemma analyzer, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[42] J. V. D. Hoven, G.-J. Lokhorst, Deontic logic and computer-supported computer ethics, Metaphilosophy 33 (3) (2002) 376–386.
URL http://www.jstor.org/stable/24439287

[43] V. Sarathy, M. Scheutz, B. Malle, Learning behavioral norms in uncertain and changing contexts, in: Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 2017.

[44] B. Liao, M. Slavkovik, L. W. N. van der Torre, Building jiminy cricket: An architecture for moral agreements among stakeholders, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019, 2019, pp. 147–153. doi:10.1145/3306618.3314257.
URL https://doi.org/10.1145/3306618.3314257

[45] R. Rzepka, K. Araki, What people say? web-based casuistry for artificial morality experiments, in: T. Everitt, B. Goertzel, A. Potapov (Eds.), Artificial General Intelligence, Springer International Publishing, Cham, 2017, pp. 178–187.

[46] F. Brandt, V. Conitzer, U. Endriss, J. Lang, A. D. Procaccia, Handbook of computational social choice, Cambridge University Press, 2016.

[47] J. Greene, F. Rossi, J. Tasioulas, K. B. Venable, B. Williams, Embedding ethical principles in collective decision support systems, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[48] V. Conitzer, W. Sinnott-Armstrong, J. S. Borg, Y. Deng, M. Kramer, Moral decision making frameworks for artificial intelligence, in: Thirty-first aaai conference on artificial intelligence, 2017.

[49] R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, A. D. Procaccia, A voting-based system for ethical decision making, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[50] R. Freedman, J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, V. Conitzer, Adapting a kidney exchange algorithm to align with human values, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[51] A. Y. Ng, S. J. Russell, et al., Algorithms for inverse reinforcement learning., in: Icml, Vol. 1, 2000, p. 2.

[52] D. Abel, J. MacGlashan, M. L. Littman, Reinforcement learning as a framework for ethical decision making, in: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[53] D. Hadfield-Menell, S. J. Russell, P. Abbeel, A. Dragan, Cooperative inverse reinforcement learning, in: Advances in neural information processing systems, 2016, pp. 3909–3917.