Piecemeal Knowledge Acquisition for Computational Normative Reasoning

Ilaria Canavotto
John Horty
icanavot@umd.edu
horty@umd.edu
Philosophy Department
University of Maryland
College Park, Maryland, USA

ABSTRACT

We present a hybrid approach to knowledge acquisition and representation for machine ethics—or more generally, *computational normative reasoning*. Building on recent research in artificial intelligence and law, our approach is modeled on the familiar practice of decision-making under precedential constraint in the common law. We first provide a formal characterization of this practice, showing how a body of normative information can be constructed in a way that is piecemeal, distributed, and responsive to particular circumstances. We then discuss two possible applications: first, a robot childminder, and second, moral judgment in a bioethical domain.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Applied computing** → **Law**.

KEYWORDS

machine ethics, computational normative reasoning, knowledge acquisition/representation, AI and Law

ACM Reference Format:
Ilaria Canavotto and John Horty. 2022. Piecemeal Knowledge Acquisition for Computational Normative Reasoning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3514094.3534182

1 INTRODUCTION

A central problem in the field of machine ethics—or more generally, *computational normative reasoning*—is the acquisition and representation of normative information in a form that allows for machine implementation.[1] This problem spans the fields of knowledge representation (KR) and machine learning (ML) in computer science, but also involves central issues in moral and legal philosophy.

There are two general approaches to the problem, with advantages and disadvantages that are by now well-known, but worth reviewing. The first is the top-down approach, according to which normative information is explicitly encoded in some symbolic formalism, often a logical language. An example of this approach can be found in the early efforts to represent legislative and regulatory information in a logic programming language [10, 40], an idea that has continued to be explored and refined, and more recently adapted to the representation of moral as well as legal norms [19, 27]. Other examples involve, for instance, the representation of the knowledge necessary for autonomous weapon systems to obey the rules of war [6], or for autonomous vehicles to engage in verifiably correct ethical reasoning [17].

This top-down approach has two central advantages. First, the meaning of the representations involved is clear, often defined by a precise semantic theory; as a result, the normative principles encoded in these representations can sensibly be challenged or justified. And second, these symbolic representations tend to support a style of computation that leads to transparent, explainable decisions—it is easy enough, for example, to understand exactly how a logic program supports the conclusions it does. The central disadvantage of the top-down approach is that it is simply not realistic to imagine that any significant body of normative information could be encoded by hand, due to the exception-laden nature of normative rules and the fact that these rules are often stated using open-textured predicates, which would require further interpretation.

Standing in contrast to the top-down approach is the bottom-up approach, according to which, in its more usual formulations, normative information is acquired through ML techniques, such as reinforcement learning or inverse reinforcement learning [1, 37] and encoded, for example, in a reward function or in a distribution of weights in a neural network. There are also a number of less usual formulations that would naturally be classified as bottom-up, such as the idea that artificial agents can learn values by reading stories [32].

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534182

[1] We use the phrase "computational normative reasoning" rather than the more common "machine ethics" in order to emphasize that the reasoning under consideration might include, not just ethical reasoning, but various other kinds of normative reasoning, such as legal or regulatory reasoning; for an example of normative reasoning that is neither ethical nor legal, see Section 5.1 below.

The central advantage of this bottom-up approach is that it avoids the knowledge acquisition bottleneck—complex normative information need not be explicitly encoded, but can be acquired implicitly, through interaction with training data. Further, the ML techniques at work in typical bottom-up systems have proved to be strikingly successful in other domains, such as pattern recognition, facial recognition, and text understanding. It is therefore not unreasonable to hope that these techniques might allow a machine to learn complex moral information as well, and at least one prominent ethicist has advocated this possibility, not on the basis of computational considerations, but instead, starting with a particular position concerning the nature of ethical theory [16].

The central disadvantage of the bottom-up approach is that, although learning may indeed take place, it is often unclear exactly what normative information has been learned: how are decisions based on this information supposed to be explained, or more important in the normative setting, justified. Consider a concrete case. Suppose a prisoner has been denied parole. This decision requires justification. But would it count as a justification to be told that an algorithm, trained on a particular data set, predicts that the prisoner will be a recidivist?

Because of the difficulties facing pure top-down or pure bottom-up approaches to the acquisition and representation of normative information, a number of researchers have begun to explore hybrid approaches, combining explicit symbolic representation with machine learning. These hybrid approaches have been developed in different domains and adapted for different reasoning tasks. For example, one early, well-known system, initially explored in the bioethical domain but then extended to several others [4, 5], represents particular decisions as vectors, with the vector components standing for the extent to which various prima facie moral principles are satisfied or violated, as a result of that decision; these decisions are classified as right or wrong by domain experts, and then the general rules thought to guide this classification arrived at through inductive logic programming. More recently, it has been suggested [41] that a particular hybrid architecture might help medical professionals make allocation decisions for organ donations. On this approach—which we will return to later for comparison—morally relevant features of potential donor recipients are first identified by domain experts; preferences over competing clusters of these features are elicited from members of a population, and on the basis of these preferences, ML techniques allow the system to offer recommendations.

The primary goal of the present paper is to present a different hybrid approach, with a distinct pattern of advantages. As with bottom-up approaches, the current approach acquires normative information from judgments in particular circumstances. But as with top-down approaches, this information is represented in symbolic form. The representation is also simple—there is no need for careful articulation of complex normative rules—and supports a natural notion of normative constraint. Finally, the approach described here is realistic, in the sense that it is based on a familiar human practice—the practice of decision making under precedential constraint in the common law. Although legal reasoning in general, and especially common law reasoning, is often viewed as obscure and contentious, the topic has recently been considerably clarified by the development of formal models within the field of artificial intelligence and law (AI and Law). The approach suggested here is based on one of these formal models [22, 24], according to which the common law constructs a body of normative information in a way that is piecemeal, distributed, and responsive to particular circumstances. A secondary goal of this paper, then, is simply to highlight the importance of research in AI and Law for computational normative reasoning more generally.

We begin in the next section with an informal description of common law reasoning and constraint. A formal characterization is set out in Sections 3 and 4, and then adapted to normative reasoning more generally and compared to a different hybrid approach in Section 5. A few open issues as well as future work are discussed in the conclusion.

2 THE COMMON LAW

The common law emerges, not from explicit legislation, but from decisions in particular cases, which then govern later cases through a complex doctrine of precedent. According to this doctrine, decisions by earlier courts constrain the decisions available to later courts while still allowing these later courts the freedom to respond to new situations in creative ways.

It is generally thought that precedential constraint is carried by rules. A precedent case normally contains, not only a factual description of some situation together with a decision on the basis of those facts, but a rule through which that decision is justified. Some writers argue that a case rule of this kind, once introduced, must then govern any later situation to which it is applicable [3]. The more standard view, however, is that later situations can be *distinguished*—where distinguishing a new situation involves identifying important, or material, differences between that situation and some earlier case in which a rule was formulated, with the result that the earlier rule is modified to avoid inappropriate application in the new situation [25, 30, 39].

This process of rule modification could be illustrated with a legal example, but it will be simpler to concentrate on a domestic scenario. Suppose, then, that Jack and Jo are the parents of two children—Emma, who has just turned nine, and Max, age twelve—and that they agree to respect each other's decisions concerning the children, treating these decisions, in effect, as precedents. And imagine that, one night, Emma, who has completed both her chores and her homework, but did not finish dinner, asks Jo if she can stay up and watch TV. This is like a legal case: a situation is presented to an authority, Jo, who must make a decision and provide a rationale for her decision. Suppose that Jo resolves the case by granting the request, stating that Emma can stay up to watch TV since she is now nine years old. This decision can be seen as introducing a household version of a common law rule—perhaps, "Children age nine or greater can stay up and watch TV"—fashioned in response to a particular set of circumstances, but applicable to future situations as well.

Now imagine that, the next day, Max, who has likewise completed chores and failed to finish dinner, but who has, in addition, failed to complete homework, asks Jack whether he can stay up and watch TV. And suppose that, in this case, Jack refuses, on the

grounds that Max has not completed his homework. Max might reasonably object, pointing out that, in the previous case of Emma, a rule was established according to which children age nine or older can stay up and watch TV. The common law, however, allows Jack to defend his decision by distinguishing the two cases, arguing that the previous rule should not apply to the new case of Max, since this new case, unlike the previous case of Emma, presents the additional feature that the child in question has not completed his homework. An effect of Jack's decision would be that the rule set out by Jo in the case of Emma is modified to avoid application in the case of Max—perhaps now understood to mean "Children age nine or greater can stay up and watch TV, unless they have failed to complete their homework."

Although this kind of normative development seems very natural, even outside the law, it leads to a dilemma concerning rule modification. On one hand, if rule modification is not allowed, it is hard to see how we could understand the process of normative development outlined in the example. But if rule modification is allowed, on the other hand, it is hard to understand the concept of precedential constraint—how can later decision makers be constrained by rules formulated in earlier decisions if they are then free to modify those rules at will?

To avoid this rule-modification dilemma, we present here an entirely different approach to precedential constraint. This new approach—which we describe as the *reason model* of constraint—takes reasons, rather than rules, as fundamental.[2] According to the reason model, what matters about an earlier decision is the court's assessment of importance among the competing reasons presented by that case, represented here as a priority ordering among these reasons. Later courts are then constrained, not to follow the rules set out in precedent cases, but simply to reach decisions that are consistent with the priority ordering that has been established earlier. Normative development of the common law is therefore pictured, not as the elaboration of an increasingly complex system of rules, but instead as the gradual construction of an increasingly rich priority ordering among reasons.

3 BASIC CONCEPTS

3.1 Factors and fact situations

For simplicity, we refer to any individual or entity with the capacity to render authoritative decisions in some domain as a *court*—for instance, Jack and Jo function as courts in their household normative system.

We suppose that a situation presented to a court for decision can be represented as a set of *factors*, where a factor is a legally significant fact or pattern of facts bearing on that decision. In our domestic scenario, the legal, or quasi-legal, issue at hand is whether a child can stay up and watch TV, and the factors involved might reasonably include those already considered—whether the child has reached the age of nine, completed chores, eaten dinner, finished homework—as well as countless others.

The factor-based representation of legal situations is not restricted to everyday examples of this kind. In fact, this style of representation has been used to analyze case-based reasoning in a number of complex legal domains within AI and Law, beginning with a careful analysis of trade-secrets law [7, 33]. In this domain, a case typically concerns the issue of whether a defendant has gained an unfair competitive advantage over a plaintiff through the misappropriation of a trade secret; and here the factors involved might turn on, say, questions concerning whether the plaintiff took measures to protect the trade secret, whether a confidential relationship existed between the plaintiff and the defendant, whether the information acquired was reverse-engineerable or in some other way publicly available, and the extent to which this information did, in fact, lead to a real competitive advantage for the defendant.

We assume that factors have polarities, favoring one of two sides of a given issue, which we refer to as π and δ, representing the plaintiff and the defendant. We take $F^\pi = \{f_1^\pi, \ldots, f_n^\pi\}$ as the set of factors favoring the plaintiff, $F^\delta = \{f_1^\delta, \ldots, f_m^\delta\}$ as the set of factors favoring the defendant, and $F = F^\pi \cup F^\delta$ as the entire set of factors. Where s is one of these sides, we let $\overline{s}$ represent the other: $\overline{\pi} = \delta$ and $\overline{\delta} = \pi$.

A *fact situation* X, of the sort presented to the court for judgment, is defined simply as some subset of factors: $X \subseteq F$. And where X is a fact situation of this kind, we let X^s represent the factors from X that support the side s, so that: $X^\pi = X \cap F^\pi$ and $X^\delta = X \cap F^\delta$. To illustrate: the situation $X_1 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$ contains two factors each favoring the plaintiff and the defendant, with $X_1^\pi = \{f_1^\pi, f_2^\pi\}$ and $X_1^\delta = \{f_1^\delta, f_2^\delta\}$.

3.2 Reasons and rules

When presented with a fact situation, a court's primary task is to reach a decision, or determine an *outcome*, where the two possible outcomes are π or δ, representing a decision for the plaintiff or defendant.

In addition to deciding for one side or the other, the court is expected to supply a rule, or principle, to justify its decision. Rules of this kind will be characterized in terms of reasons, where a *reason for a side* is some set of factors uniformly favoring that side; a *reason* can then be defined as a set of factors uniformly favoring one side or another. To illustrate: $\{f_1^\pi, f_2^\pi\}$ is a reason favoring the plaintiff, and so a reason, while $\{f_1^\delta, f_2^\delta\}$ is a reason favoring the defendant.

We stipulate that a reason U *holds* in a situation X just in case $U \subseteq X$. And we define a relation of strength among reasons for a side according to which, where U and V are reasons for the same side, then V *is at least as strong a reason as* U *for that side* just in case $U \subseteq V$. To illustrate: the reason $\{f_1^\pi\}$ holds in the previous fact situation $X_1 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$, since $\{f_1^\pi\} \subseteq X_1$ and of the two reasons $\{f_1^\pi\}$ and $\{f_1^\pi, f_2^\pi\}$, the second favors the plaintiff at least as strongly as the first, since $\{f_1^\pi\} \subseteq \{f_1^\pi, f_2^\pi\}$.

Given this notion of a reason, a *rule* can now be defined as a statement of the form $U \rightarrow s$, where U is a reason supporting the side s. For convenience, we introduce two functions—*Premise* and *Conclusion*—picking out the premise and conclusion of a rule. To illustrate: the statement $\{f_1^\pi\} \rightarrow \pi$ is a rule, since $\{f_1^\pi\}$ is a reason supporting the plaintiff. If we take r_1 to stand for this rule, we have $Premise(r_1) = \{f_1^\pi\}$ and $Conclusion(r_1) = \pi$.

[2] This shift in emphasis from rules to reasons reflects a general theme in recent work in ethics and normative theory [43].

The rules defined here are to be interpreted as defeasible, telling us that their premises entail their conclusions, not as a matter of necessity, but only by default. What the rule $r_1 = \{f_1^\pi\} \rightarrow \pi$ means, very roughly, is that, whenever the premise $\{f_1^\pi\}$ of the rule holds in some situation, then, as a default, the court ought to decide that situation for the conclusion π of the rule—or perhaps more simply, that the premise of the rule provides the court with a reason for deciding in favor of its conclusion.

3.3 Cases and case bases

A *case* can now be defined as a situation together with an outcome and a rule through which that outcome is justified: such a case can be specified as a triple of the form $c = \langle X, r, s\rangle$, where X is a situation containing the factors presented to the court, r is a rule, and s is an outcome. We refer to r as the *rule of the case*, and to $Premise(r)$, the reason that forms the premise of this rule, as the *reason for the decision* in that case—and since reasons and rules are so closely related, we will say, indifferently, that the case is decided *on the basis of* either the rule or the reason that forms its premise.

We introduce three more auxiliary functions—*Facts*, *Rule*, and *Outcome*—mapping cases into their component parts, so that, in the case c above, we would have $Facts(c) = X$, $Rule(c) = r$, and $Outcome(c) = s$. And in order for the concept of a case to make sense, we stipulate that the premise of a case rule must hold in the fact situation of the case, and that the rule's conclusion must match the case outcome—or that $Premise(r) \subseteq Facts(c)$ and $Conclusion(r) = Outcome(c)$.

This concept can be illustrated with the case $c_1 = \langle X_1, r_1, s_1\rangle$, where the fact situation of this case is the familiar $X_1 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$, where the case rule is the familiar $r_1 = \{f_1^\pi\} \rightarrow \pi$, and where the outcome of the case is $s_1 = \pi$, a decision for the plaintiff. This case represents a situation in which the court is confronted with the fact situation X_1, decided for the plaintiff on the basis of the rule r_1, according to which the presence of the factor f_1^π—that is, the reason $\{f_1^\pi\}$—leads, by default, to a decision for the plaintiff.

Finally, a *case base* is defined as a set Γ of cases. It is a case base of this sort—a set of precedent cases—that will be taken to represent the common law in some area, and to constrain the decisions of future courts.

4 CONSTRAINT BY REASONS

According to the reason model, later courts are constrained to reach decisions that are consistent with the priority ordering among reasons derived from decisions of earlier courts. In order to develop this suggestion, we need to explain how a priority ordering on reasons can be derived from the decisions of earlier courts, and then what it means for the decision of a later court to be consistent with that ordering.

4.1 A priority ordering on reasons

To begin with, let us return to the case $c_1 = \langle X_1, r_1, s_1\rangle$—where $X_1 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$, where $r_1 = \{f_1^\pi\} \rightarrow \pi$, and where $s_1 = \pi$—and ask what information is carried by this case; what is the court telling us with its decision? Well, two things.

First, by deciding for the plaintiff on the basis of the rule r_1, the court is registering its judgment that $Premise(r_1)$, the reason for its decision, is more important—or has higher priority—than any reason for the defendant that holds in X_1, the fact situation of the case. How do we know this? Because if the court thought some reason for the defendant that held in the situation X_1 was more important than $Premise(r_1)$, the court would have found for the defendant on the basis of that reason, rather than for the plaintiff on the basis of $Premise(r_1)$.

And second, if the court is telling us explicitly that the reason $Premise(r_1)$ itself is more important than any reason for the defendant that holds in X_1, then the court must also be telling us, at least implicitly, that any other reason for the plaintiff that is at least as strong as $Premise(r_1)$ must likewise be more important than any reason for the defendant that holds in this situation.

A reason U for the defendant holds in the situation X_1 just in case $U \subseteq X_1$, and a reason V for the plaintiff is at least as strong for the plaintiff as the reason $Premise(r_1)$ just in case $Premise(r_1) \subseteq V$. If we let the relation $<_{c_1}$ represent the priority ordering on reasons derived from the particular case c_1, then, the force of the court's decision in this case is simply that: where U is a reason favoring the defendant and V is a reason favoring the plaintiff, then $U <_{c_1} V$ just in case $U \subseteq X_1$ and $Premise(r_1) \subseteq V$. Consider, for example, the reason $\{f_1^\delta\}$ for the defendant and the reason $\{f_1^\pi, f_2^\pi, f_3^\pi\}$ for the plaintiff. Here, we have $\{f_1^\delta\} \subseteq X_1$ as well as $Premise(r_1) \subseteq \{f_1^\pi, f_2^\pi, f_3^\pi\}$. We therefore have $\{f_1^\delta\} <_{c_1} \{f_1^\pi, f_2^\pi, f_3^\pi\}$—the court's decision in c_1 entails that the reason $\{f_1^\pi, f_2^\pi, f_3^\pi\}$ favoring the plaintiff is to be assigned a higher priority than the reason $\{f_1^\delta\}$ favoring the defendant.

Generalizing from this example, we reach the following definition of the priority ordering among reasons derived from a single case:

Definition 1 (Priority ordering derived from a case). Let $c = \langle X, r, s\rangle$ be a case, and let U and V be reasons favoring the sides $\overline{s}$ and s respectively. Then the relation $<_c$ representing the priority ordering on reasons derived from the case c is defined by stipulating that $U <_c V$ if and only if $U \subseteq X$ and $Premise(r) \subseteq V$.

Once we have defined the priority ordering on reasons derived from a single case, we can introduce a priority ordering $<_\Gamma$ derived from an entire case base Γ by stipulating that one reason has a higher priority than another according to the case base whenever that priority is supported by some particular case from the case base:

Definition 2 (Priority ordering derived from a case base). Let Γ be a case base, and let U and V be reasons. Then the relation $<_\Gamma$ representing the priority ordering on reasons derived from the case base Γ is defined by stipulating that $U <_\Gamma V$ if and only if $U <_c V$ for some case c from Γ.

And using this concept of the priority ordering derived from a case base, we can now define a case base itself as inconsistent if its derived ordering yields conflicting information about the priority among reasons—telling us, for some pair of reasons, that each has a higher priority than the other—and consistent otherwise:

Definition 3 (Inconsistent and consistent case bases). Let Γ be a case base with $<_\Gamma$ its derived priority ordering. Then Γ is inconsistent if and only if there are reasons U and V such that $U <_\Gamma V$ and $V <_\Gamma U$, and consistent otherwise.

4.2 Constraint

We now define a notion of constraint according to which a court confronted with a new situation X against the background of a consistent case base Γ is required simply to reach a decision in X that preserves the consistency of Γ—that is, a decision that does not introduce inconsistency into the case base. Our account applies, in the first instance, to the rules on the basis of which a court is permitted to justify its decisions:

Definition 4 (Constraint on rule selection). Let Γ be a consistent case base and X a fact situation confronting the court. Then the court is permitted to base its decision in X on some rule r supporting an outcome s such that the augmented case base $\Gamma \cup \{\langle X, r, s\rangle\}$ remains consistent.

This idea can be illustrated by assuming as background the case base $\Gamma_1 = \{c_1\}$, containing as its single member the familiar case $c_1 = \langle X_1, r_1, s_1\rangle$—where, again, $X_1 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$, where $r_1 = \{f_1^\pi\} \rightarrow \pi$, and where $s_1 = \pi$. Suppose that, against this background, the court confronts the fresh situation $X_2 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta, f_3^\delta\}$ and considers finding for the defendant in this situation on the basis of the reason $\{f_1^\delta, f_2^\delta\}$, leading to the decision $c_2 = \langle X_2, r_2, s_2\rangle$, where X_2 is as above, where $r_2 = \{f_1^\delta, f_2^\delta\} \rightarrow \delta$, and where $s_2 = \delta$. Is the court permitted to carry through with this plan of action?

Well, as we can see, $Premise(r_1) = \{f_1^\pi\}$, the reason for the decision in the initial case, holds in the new situation X_2 as well, since $\{f_1^\pi\} \subseteq X_2$. And of course, the new reason $Premise(r_2) = \{f_1^\delta, f_2^\delta\}$ favors the defendant at least as strongly as itself—that is, $Premise(r_2) \subseteq Premise(r_2)$, or $Premise(r_2) \subseteq \{f_1^\delta, f_2^\delta\}$. It therefore follows from Definition 1 that c_2, the court's envisaged decision, would assign the reason $\{f_1^\delta, f_2^\delta\}$ for the defendant a higher priority than the reason $\{f_1^\pi\}$ for the plaintiff—that is, $\{f_1^\pi\} <_{c_2} \{f_1^\delta, f_2^\delta\}$. But Γ_1 already contains the case c_1, from which, in a similar fashion, we can derive the priority relation $\{f_1^\delta, f_2^\delta\} <_{c_1} \{f_1^\pi\}$, telling us exactly the opposite. Since the augmented case base

$$\begin{aligned}\Gamma_2 &= \Gamma_1 \cup \{c_2\}\\ &= \{c_1, c_2\}\end{aligned}$$

resulting from the court's envisaged decision contains both these cases, we would then have both $\{f_1^\delta, f_2^\delta\} <_{\Gamma_2} \{f_1^\pi\}$ and $\{f_1^\pi\} <_{\Gamma_2} \{f_1^\delta, f_2^\delta\}$ by Definition 2, so that, by Definition 3, this augmented case base would be inconsistent. By Definition 4, then, we can conclude that the court is not permitted to carry through with its plan of deciding for the defendant in the situation X_2 on the basis of the rule r_2, since c_2, the resulting decision, would introduce an inconsistency into the background case base, but the reason model requires decisions to preserve case base consistency.

Of course, it does not follow from the fact that the court is not permitted to decide for the defendant in the situation $X_2 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta, f_3^\delta\}$ on the basis of the particular rule r_2 that it cannot decide for the defendant in this situation at all. Suppose, for example, that the court appeals to the reason $\{f_1^\delta, f_3^\delta\}$ to justify its decision for the defendant, leading to the case $c_3 = \langle X_3, r_3, s_3\rangle$, where $X_3 = X_2$, where $r_3 = \{f_1^\delta, f_3^\delta\} \rightarrow \delta$, and where $s_3 = \delta$. The augmented case base

$$\begin{aligned}\Gamma_3 &= \Gamma_1 \cup \{c_3\}\\ &= \{c_1, c_3\}\end{aligned}$$

resulting from this decision would then be consistent. As before, the previous case c_1 supports the priority $\{f_1^\delta, f_2^\delta\} <_{c_1} \{f_1^\pi\}$, and the new decision c_3 would now support the priority $\{f_1^\pi\} <_{c_3} \{f_1^\delta, f_3^\delta\}$, so that we would then have both the case base priorities $\{f_1^\delta, f_2^\delta\} <_{\Gamma_3} \{f_1^\pi\}$ and $\{f_1^\pi\} <_{\Gamma_3} \{f_1^\delta, f_3^\delta\}$. But there is nothing inconsistent about this pair of priorities, as we can see, informally at least, with another domestic example: one can easily imagine a teenager thinking, and thinking consistently, that going to the movies is more fun than going to the beach with her parents, but that going to the beach with her friends is more fun than going to the movies.

Our suggestion is that this is how the common law develops in the normal, incremental case—by building up a stronger and stronger priority ordering on reasons in a piecemeal fashion, through a series of decisions that are, at each stage, consistent with the existing case base.

4.3 The domestic scenario

All of this has been very abstract. For a more concrete illustration of the reason model we return to the domestic scenario example set earlier, in Section 2, concerning the questions presented by Max and Emma to their parents, Jack and Jo. As we recall, Emma is nine, failed to finish dinner, but completed homework; Max is twelve and neither finished dinner nor completed homework. Both children wanted to stay up and watch TV. In our scenario, Emma first asked for permission from Jo, who granted the request on the grounds that Emma was at least nine years old. Max then asked for permission from Jack, who denied the request—even though Max too was at least nine—on the grounds that Max had failed to complete homework.

With the children as plaintiffs and the parents as both defendants and adjudicators, or courts, this example can be cast in our framework by letting the factor f_1^π represent the fact that the child in question is at least nine years old, f_2^π the fact that the child in question completed chores, and then f_1^δ and f_2^δ, respectively, the facts that the child failed to finish dinner and failed to complete homework. The initial situation presented by Emma to Jo can be represented as $X_4 = \{f_1^\pi, f_2^\pi, f_1^\delta\}$, which Jo then decided for Emma on the basis of the rule $r_4 = \{f_1^\pi\} \rightarrow \pi$, leading to the decision $c_4 = \langle X_4, r_4, s_4\rangle$, where X_4 and r_4 are as above, and where $s_4 = \pi$. As a result of this initial decision, the case base representing the common law of the household, at least as it pertains to staying up and watching TV, is $\Gamma_4 = \{c_4\}$, with $<_{\Gamma_4}$ as its associated ordering on reasons.

Next, the situation presented by Max to Jack can be represented as $X_5 = \{f_1^\pi, f_2^\pi, f_1^\delta, f_2^\delta\}$. In keeping with our story, we suppose that Jack would like to decide against Max on the basis of the rule

$r_5 = \{f_2^\delta\} \rightarrow \delta$, leading to the decision $c_5 = \langle X_5, r_5, s_5\rangle$, where X_5 and r_5 are as above, and where $s_5 = \delta$. Is he permitted to do so, against the background of the case base Γ_4?

The answer is Yes. From Jo's earlier decision, we can conclude that the reason $\{f_1^\pi\}$ is assigned a higher priority than the reason $\{f_1^\delta\}$—that $\{f_1^\delta\} <_{c_4} \{f_1^\pi\}$, so that $\{f_1^\delta\} <_{\Gamma_4} \{f_1^\pi\}$ as well. And Jack's decision would force us to conclude also that the reason $\{f_2^\delta\}$ must be assigned a higher priority than the reason $\{f_1^\pi\}$—that $\{f_1^\pi\} <_{c_5} \{f_2^\delta\}$. But there is no conflict between this priority statement and the previous priority statement, derived from Jo's decision—a reasonable individual might, for example, prefer chocolate ice cream to vanilla and vanilla to strawberry. And because the background case base Γ_4 currently contains only Jo's decision, it follows that Jack's decision in the case of Max is consistent with this case base as well. As a result of this decision, the augmented case base

$$\begin{aligned}\Gamma_5 &= \Gamma_4 \cup \{c_5\}\\ &= \{c_4, c_5\}\end{aligned}$$

now represents the household normative system, with $<_{\Gamma_5}$ as its strengthened ordering on reasons.

4.4 Requirements and permissions

The notion of constraint set out in Definition 4 characterizes the rules on the basis of which a court is permitted to base its decisions. But of course, once this notion is in place, it can be used to define the decisions that a court is permitted, or required, to make—through the stipulation that a court is permitted to decide for a particular side if some permitted rule supports that side, and required to decide for a side if every permitted rule supports that side.

To make this clear, we let the statement $s(X)$ mean that the situation X is decided for the side s, so that—taking $\bigcirc$ and P as the usual requirement and permission operators from deontic logic—the statements $\bigcirc s(X)$ and $\mathrm{P}s(X)$ stand for the statements that the court is, respectively, required or permitted to decide X for the side s. These ideas are defined as follows:

Definition 5 (Constraint on decision). Let Γ be a consistent case base and X a fact situation confronting the court. Then it follows from Γ that the court is required to reach a decision in X for the side s—that is, $\bigcirc s(X)$—if and only if every rule on the basis of which the court is permitted to decide X supports s; the court is permitted to reach a decision in X for s—that is, $\mathrm{P}s(X)$—if and only if some rule on the basis of which the court is permitted to decide X supports s.

For illustration, we return to our domestic scenario. Here, as we have seen, it follows from Γ_4 that $\mathrm{P}\delta(X_5)$—Jack is permitted to decide the situation presented by Max in favor of the defendant. The scenario developed this far presents no interesting requirements, but suppose Jack and Jo have another child, Lynn, who would also like to stay up and watch TV, and of whom it is known only that she is older than nine but failed to finish dinner, so that the situation she presents is $X_6 = \{f_1^\pi, f_1^\delta\}$. We can then verify that $\bigcirc\pi(X_6)$ follows from the case base Γ_4, so that, if Jack confronts this new situation against the background of Jo's decision in the previous case of Emma, he is required to decide for Lynn, the plaintiff.

The operators introduced here can be shown to define a simple, and sensible, deontic logic. For example, the statement $\bigcirc s(X)$ follows from some case base just in case $\mathrm{P}\bar{s}(X)$ does not—the court is required to decided X for s just in case it is not permitted to decide X for $\bar{s}$, the opposite side. Further, as long as a case base is consistent, it will never support both $\bigcirc s(X)$ and $\bigcirc\bar{s}(X)$—the court will never be required to decide the same situation for one side and also for the other side. Finally, exactly one of the formulas $\bigcirc(X)$ or $\bigcirc\bar{s}(X)$ or $\mathrm{P}s(X) \wedge \mathrm{P}\bar{s}(X)$ will always be supported—so that, in any situation, a decision is either required for one side, or required for the other side, or it is permissible to decide for either side.

Deontic logics validating properties like these are frequently appealed to in the top-down approach to computational normative reasoning, either for system specification [6] or in the design of reasoning engines [12]. As noted earlier, the problem presented by this top-down approach centers around acquisition of the knowledge encoded in these logics. The current framework suggests one solution to this problem, according to which this knowledge is derived from particular decisions in concrete cases. In a very real sense, a system designed in accord with the current approach can be said to learn the appropriate deontic principles from these particular decisions, just as a legal system can be said to learn the common law from the particular decisions of individual courts.

5 APPLICATIONS

We discuss two possible applications of the approach presented here—one is hypothetical at this point, but still worth thinking about; the other is less hypothetical.

5.1 A robot childminder

Our hypothetical example builds on the domestic scenario already considered. Suppose that, one night a week, Jack and Jo leave their children—Emma, Max, and now Lynn—with Charlie, a robot childminder. Besides entertaining the children, Charlie's main task is to tell Emma and Max when to go to bed, and perhaps to call Jack and Jo in case the children do not follow instructions. From a design perspective, this task raises an interesting question: How are the parents able to communicate the appropriate bedtime to Charlie—how does Charlie learn when the children are supposed to go to bed?

One answer might be that Charlie has a bedtime parameter that Jack and Jo can simply set. So, imagine that, after purchasing Charlie, Jack and Jo ask each other how they should set this parameter. A discussion ensues:

> *Normally*, the children should go to bed at 9:00pm. *But*, if they are sufficiently good during the evening—that is, they complete their chores, they complete their homework, they finish dinner, and so on—and they ask to stay up and watch TV, then they can stay up until 9:30pm, *unless* they have a school trip planned for the next day …

Perhaps anticipating a discussion like this, the designers of Charlie might decide that it is better to supply the robot with a rule-based reasoning module, so that parents can specify bedtime, not by setting a parameter, but by formulating the appropriate rules. So now

imagine that Jack and Jo begin with a rule like: "Children should go to bed at 9:00pm unless they ask to stay up to watch TV and have completed homework, completed chores, finished dinner, and there is no school trip planned for the next day." At this point, however, the parents might realize that other factors—perhaps more distant, but certainly not out of the question—could intervene. For example, Jack and Jo might agree that, if the children fail to call their great aunt Olive to wish her a happy birthday, as they had promised to do, then they should both go to bed at 8:30pm.

Like the parameter-based approach, the simple rule-based approach to defining appropriate bedtime fails as well, and for the same reason that top-down approaches to computational normative reasoning are generally problematic: the list of exceptions to any given normative rule is open-ended, and cannot be anticipated in advance.

The reason model presented here suggests an alternative, more promising architecture for Charlie. The idea is that, instead of relying on a single bedtime parameter to be fixed in advance, or a set of rules defining the appropriate bedtime, Charlie learns the household concept of bedtime from Jack's and Jo's particular decisions in concrete cases. A central component of this alternative architecture is a memory of past decisions. Assuming that Charlie represents a case as a fact situation together with an outcome and a rule through which that outcome is justified, Charlie's memory of these past decisions works as a case base. In particular, as explained in Section 4, it can be used to define an ordering on reasons that determines which decisions Charlie will be required or permitted to make in future situations. We now briefly consider this general architecture and a few of the issues raised by it.

Initial training period: Suppose that Jack and Jo turn on Charlie for the first time. Before being left alone with the children, the robot needs to be provided with examples of concrete bedtime decisions—a set of training examples, constituting its initial memory. Charlie could acquire this set of training examples in at least two ways. It might ask the parents to complete a questionnaire stating their decisions in a number of paradigmatic bedtime-scenarios, and identifying reasons for those decisions. Or, in a more sophisticated version, Charlie might follow the parents around the house during a specified training period, observing bedtime decisions and querying them for reasons, in case the reasons behind their decisions are not obvious.

This suggested procedure for initial training raises a number of issues, of which we mention two. First, what happens if, during the training period, Charlie observes the parents making inconsistent decisions. According to the reason model, Charlie cannot simply incorporate these inconsistent decisions into its memory, or case base, since the reason model notion of constraint, set out in Section 4, requires decisions to preserve consistency of the underlying case base; this, of course, presupposes that the case base is consistent to start with.

One way around this problem might be to imagine that, when the parents are observed making inconsistent decisions during the training phase, Charlie informs them of the inconsistency, explains the decisions that generated it, asks them to revise one of these decisions, and adds the revised decision to its memory. Charlie could then be seen as performing, in addition to its primary function as a childminder, an additional function as a consistency checker for parental decisions—this additional function may be useful in itself since, as everyone knows, children are finely-tuned to detect and exploit any hint of inconsistency in parental decisions!

But even if it might be feasible for a short training period, asking parents to maintain strict consistency in their decisions over the longer term is not realistic. Fortunately—although we cannot discuss this in any detail here—it is possible to generalize the reason model notion of constraint to apply to inconsistent case bases as well; the key idea is that decisions should be required, not necessarily to preserve consistency of a consistent case base, but simply to introduce no new inconsistencies into a case base that may already be inconsistent.[3]

A different issue arises when we ask how many of the parents Charlie is monitoring during the training period. Suppose Charlie is monitoring only one parent—say, Jo. Then as it develops its memory, or case base, Charlie can be seen as learning Jo's preferred household normative system, at least as it bears on bedtime decisions. But suppose Charlie is monitoring both parents—Jack and Jo—during the training period and that the two parents, while respecting each other's decisions, work with slightly different bedtime standards. In that case, Charlie is learning neither Jo's nor Jack's normative system, exactly, but instead, a system that combines the two parents' normative views in a particular way—just as the common law proper, evolving over time in responses to different fact situations presented to different courts, does not necessarily reflect the views of any one court, or any one segment of society.

Charlie at work: Once the training period is complete and Charlie's initial memory contains enough bedtime decisions, the robot will then be able to reach reasonable decisions in future situations concerning what the children are required or permitted to do. For example, suppose that, during training, Charlie observed Jo allowing Emma—age nine, who failed to finish dinner, but completed homework—to stay up and watch TV on the grounds that she is now at least nine years old. Charlie's memory, or case base, will then contain the case c_4, defined in Section 4.3, representing this decision. Imagine that Charlie next confronts the situation X_6 presented by Lynn—also nine, who has likewise failed to finish dinner. Then as we saw in Section 4.4, Charlie will conclude $\bigcirc\pi(X_6)$—the robot will be required to reach a decision in favor of Lynn, the plaintiff.

If Charlie's memory is rich enough, most new situations will be settled in this way—Charlie will be required to reach a decision for one side, or the other. But there may still be situations in which Charlie is permitted to decide either way. Suppose, for example, that Charlie's memory, or case base, contains only c_4, representing Jo's decision concerning Emma, and that the robot then confronts the situation X_5 presented by Max—age twelve, who neither finished dinner nor completed chores. We saw in Section 4.4 that a case base like this supports $\mathsf{P}\delta(X_5)$, and it is easy enough to see that the same case base supports $\mathsf{P}\pi(X_5)$ as well—the result is that Charlie is permitted to decide this situation for the defendant or the plaintiff, against Max or in favor of Max.

[3] See Section 2.2.2 of [23] for a discussion of this generalization. A different approach to the problems presented by inconsistent case bases can be found in [44].

In such a situation, where either decision is permitted, it seems sensible that Charlie should be required to call Jack and Jo, explain the situation, let them decide, inform Max of their choice, and update its memory, or case base, with this new decision. It may be, however, that Jack and Jo—who are, after all, trying to enjoy a night away from their children—become tired of Charlie's queries. In that case, since both decisions are in fact permitted, they may authorize Charlie simply to choose the one it thinks is best, perhaps drawing on other computational resources to make that decision. The parents could then later review Charlie's decision and its rationale before adding that new decision to the robot's memory, so that, until review, Charlie's independent decision plays no role in determining what is required or permitted in the future.

5.2 Kidney allocation decisions

The second application we consider—less hypothetical—starts with the idea of deploying ML techniques to aid human moral judgment in the bioethical domain. This idea is developed, in particular, in a number of papers [15, 18, 42] centered around moral aspects of kidney allocation decisions; the work is presented from a philosophical standpoint in [41].

Setting aside multi-party kidney exchanges [36], we follow [41] in considering only a simple kidney allocation decision: a single kidney is available for transplant but there are two potential recipients—Patient A and Patient B—both of whom need a kidney. Which one of the two should get the single available kidney? A decision like this is often constrained by medical factors, such as blood type compatibility or organ quality. But when medical factors do not settle the issue, many individuals believe that moral features of the potential recipients should play a role in organ allocation decisions; these features might include, for example, number of dependents of the potential recipients, history of violent crime, or whether the disease was caused by alcoholism or drug use.

Whether or not moral features like these should, in fact, be taken into account in kidney allocation decisions is a contentious issue in bioethics, which we do not consider here. But even supposing it is appropriate to consider these features, we are still faced with two questions. First, which moral features, exactly, should we consider? And second, how do we evaluate the relative importance of these features—or more generally, once two potential kidney recipients are described in terms of their morally relevant features, how do we use this information to decide which one gets the single available kidney?

The approach outlined in [41] sketches an answer to both questions. The relevant features are identified through a complex process that involves: first, crowdsourcing; second, surveys of domain experts, such as doctors and hospital administrators; third, consideration of those features already identified as important in the ethical literature; fourth, editing the preliminary feature list for clarity, redundancy, and completeness; and fifth, validation of the feature list through further testing. The authors imagine this process of refinement and validation might have to be repeated several times, but that it will eventually result in a partition of features into those that are morally relevant, those that are morally irrelevant, and those whose moral relevance is controversial.

Turning to the second question, the authors of [41] pursue a ML approach. They begin by introducing the notion of a *conflict*, defined as a pair of potential kidney recipients characterized in terms of the features already identified as morally relevant. To illustrate, the pair consisting of Patient A and Patient B, characterized in terms of the indicated features, would constitute a conflict:

Patient A
36 years old
0 child dependents
3 drinks per day prediagnosis

Patient B
53 years old
1 child dependents
2 drinks per day prediagnosis

Conflicts like these are generated and presented to subjects through a web site.[4] Subjects are asked to resolve the conflicts by choosing, based on the characterizations of the two potential recipients, which one should get a single available kidney.

Let us define a *resolved conflict* as a conflict together with a subject's choice determining which potential recipient is to receive the kidney. By presenting conflicts to subjects and tracking their resolutions, the site assembles a set of resolved conflicts. Once it has gathered enough information, this set of resolved conflicts can function as a rich body of labeled training data—where the data are the conflicts presented to subjects and their labels are the resolutions to these conflicts provided by those subjects. The hope is that ML techniques can then be applied to this training data to learn enough about how features support decisions—the relative importance of the different features, how they interact—that this knowledge can then be projected forward to resolve future conflicts in a reasonable way, determining the morally preferred kidney recipient in conflicts that have not yet been considered.

The authors of [41] describe their proposal as a hybrid approach, since the set of morally relevant features is first constructed by hand, in a careful, top-down fashion, but the way in which these features interact to yield overall judgments is then arrived at through bottom-up ML techniques.

This approach is sensible and promising. We want to suggest, however, that an alternative approach, also sensible, can be developed based on the reason model of constraint described in this paper. This suggestion is motivated by a strong analogy between the problem representation and methodology outlined in [41] and the formal framework set out in this paper. To spell it out: First, the two possible kidney recipients, Patient A and Patient B, are analogous to the plaintiff and the defendant. Second, the features in the kidney domain are analogous to our factors, normatively relevant facts or patterns of facts. Third, the conflicts in the kidney domain are analogous to our fact situations, sets of features, or factors. Fourth, the subjects in the kidney domain are analogous to our courts, individuals or authoritative bodies rendering judgments for one side or the other in the face of these conflicting features, or factors. Fifth, the resolved conflicts from the kidney domain are analogous to our cases, sets of features, or factors, supplemented

[4]The site is *whogetsthekidney.com*; in fact, the particular conflict displayed in the text was generated by this site.

with the decision for one side or the other arrived at by a subject, or a court. Sixth and finally, the set of resolved conflicts from the kidney domain is analogous to the case base defined here.

The most striking disanalogy between the approach outlined in the kidney domain work and that suggested by the reason model lies in the way normative information is projected forward from a body of settled decisions. According to [41], as we have seen, ML techniques are supposed to learn a rudimentary moral theory from the set of resolved conflicts, which can then be applied to future conflicts. According to the reason model, by contrast, future courts are required only to reach decisions in future fact situations in a way that preserves consistency of the underlying case base.

This central difference should not be taken as a disagreement, suggesting that we must choose one approach over the other, but instead, as opening up a range of interesting questions—both technical and philosophical—concerning the relations between the approaches. As an example of a technical question, suppose that, working in the framework of [41], the subjects presented with conflicts resolve these conflicts in a way that is consistent, in the sense defined by the reason model, so that the resulting set of resolved conflicts is consistent as well. Then the question arises: Will the moral theory learned by an ML system that is trained on the basis of this consistent set of resolved conflicts yield decisions in future cases that are likewise consistent? Either answer would be interesting.

As an example of a philosophical question raised by these contrasting approaches, we can note that the techniques employed in [41] are similar to the techniques often used for preference aggregation—indeed, much of the literature on moral reasoning in AI draws directly on the preference aggregation literature [21, 35]. But it is also sensible to ask: How closely can the process of resolving moral differences be assimilated to the process of aggregating conflicting preferences—are there situations in which this assimilation is not plausible? Certainly the techniques explored in AI for preference aggregation do not correspond to the way in which conflicting legal decisions are resolved, for example. And even in contractualist moral systems, which view morality as arising out of a kind of negotiation between individuals with conflicting preferences [20, 38], the account of negotiation involved is much more complex than the forms of preference aggregation familiar in AI.

6 CONCLUSION

The goal of this paper has been to present a new hybrid approach to knowledge acquisition and representation for computational normative reasoning. Like the usual bottom-up approaches, the approach presented here is grounded in the judgments of individuals in concrete circumstances, rather than in complex, abstract rules formulated in advance, thus, to some extent, avoiding the knowledge-acquisition bottleneck. Like the usual top-down approaches, the current proposal represents normative information in a logical language, rather than in a reward function or a pattern of weights in a neural network, thus allowing for explainability and explicit justification of decisions. Finally, the current approach is modeled on the familiar human practice of the common law, which constructs a body of normative information in a way that is piecemeal, distributed, and responsive to particular circumstances.

The approach sketched here is implementable in a direct fashion that does not require interpretation of the formalism into any logic; the crucial fact that makes this implementation feasible is that checking consistency of a new decision against a background case base is linear in the size of the case base. On the other hand, the resulting information can be encoded into a standard defeasible logic—such as prioritized variants of default logic [31] or logics of structured argumentation [26]—to support the generation of more helpful explanations. Interpreting the normative information acquired through the approach sketched here also allows for richer, more sophisticated analyses of the normative reasoning underlying a particular decision, such as the ability to reason, not just with defeasible normative principles, but also about these principles—perhaps some principles require that others should be taken out of consideration, or excluded, for example [29].

We close by simply mentioning two open issues:

First, our approach, based on the reason model of constraint, depends on factors—normatively significant facts or patterns of facts—but where do these factors come from? As we have seen, the identification of morally relevant features in the kidney allocation domain involves a complex procedure of crowdsourcing, research, testing, and refinement. In the traditional AI and Law domain, the identification of legally relevant factors follows an equally complex knowledge engineering methodology [2]. But there are other domains in which the identification of appropriate factors seems to be relatively straightforward, such as the domestic situations considered here or some of the more ordinary cases of everyday risk [15]; in addition, ML techniques have more recently been employed [9, 13] for factor identification in the legal domain. This variety of approaches leads to the practical question: Is there anything we can say about the appropriate means of factor identification? And behind this practical question lurks a deeper philosophical question: What does it mean to identify some fact as a factor—that is, not just as an aspect of the world, but as an aspect with normative bearing on some particular issue? We are currently exploring the idea, previously hinted at in [11, 28], that what gives a mere fact normative significance is that reaching a decision on the basis of that fact promotes a value.

A different issue concerns the granularity of factors. The factors at work in the approach set out here are relatively narrow, or fine-grained—factors such as, in the domestic domain, whether a child finished homework or is now nine years old; or in the organ allocation domain, whether a patient in kidney failure habitually consumed two, or three, drinks per day prediagnosis; or in the trade-secrets domain, whether a particular design was reverse-engineerable with a given level of expertise. The granularity of these factors allows normative judgments to be responsive to subtle differences between situations; further, their application in a given situation is not controversial. On the other hand, although the current symbolic approach does allow for justification of decisions, it may turn out that the resulting justifications, formulated in terms of such low-level factors, are not particularly helpful, or satisfying—why exactly should the fact that a child is now nine years old, for example, count as a justification for a decision to allow her to stay up and watch TV?

The more traditional ethical literature, by contrast, is organized around much higher-level concepts, or factors—such as, to draw on one traditional theory, fidelity, reparation, gratitude, justice, and beneficence [34]; or to draw on another, autonomy and non-maleficence [8]. Justifications cast in terms of predicates like these would be more satisfying from a normative perspective. But it would be harder to register fine-grained moral distinctions between situations described at such a high level of generality; further, it may not be obvious whether or not these high-level concepts are applicable in a given situation—whether, for example, a particular decision exhibits sufficient respect for autonomy. What is needed is a hierarchy of normative predicates linking the kind of very low-level factors considered here with the higher-level normative concepts that give these factors their force—telling us, for example, that respect for autonomy should increase with a child's age, so that having reached the age of nine favors allowing a child to stay up and watch TV. Working out the logic of this hierarchy is likewise a topic of current research [14].

REFERENCES

[1] David Abel, James MacGlashan, and Michael Littman. 2016. Reinforcement learning as a framework for ethical decision making. In *AI, Ethics, and Society: Papers from the 2016 AAAI Workshop*, Blai Bonet, Sven Koenig, Benjamin Kuipers, Illah Nourbakhsh, Stuart Russell, Moshe Vardi, and Toby Walsh (Eds.). AAAI Press.

[2] Vincent Aleven. 1997. *Teaching Case-Based Argumentation Through a Model and Examples.* Ph. D. Dissertation. Intelligent Systems Program, University of Pittsburgh.

[3] Larry Alexander and Emily Sherwin. 2008. *Demystifying Legal Reasoning.* Cambridge University Press.

[4] Michael Anderson and Susan Leigh Anderson. 2007. Machine ethics: creating an ethical intelligent agent. *AI Magazine* 28 (2007), 15–26.

[5] Michael Anderson and Susan Leigh Anderson. 2018. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics* 9 (2018), 337–357.

[6] Ronald Arkin. 2009. *Governing Lethal Behavior in Autonomous Robots.* Chapman and Hall/CRC.

[7] Kevin Ashley. 1990. *Modeling Legal Argument: Reasoning with Cases and Hypotheticals.* The MIT Press.

[8] Tom Beauchamp and James Childress. 1985. *Principles of Biomedical Ethics.* Oxford University Press.

[9] Trevor Bench-Capon and Katie Atkinson. 2021. Precedential constraint: the role of issues. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL 2021).* The Association for Computing Machinery Press, 12–21.

[10] Trevor Bench-Capon, Gwen Robinson, Tom Routen, and Marek Sergot. 1987. Logic programming for large scale applications in law: a formalisation of supplementary Benefit Legislation. In *Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL-87).* The Association for Computing Machinery Press, 190–198.

[11] Trevor Bench-Capon and Giovanni Sartor. 2001. Theory based explanation of case law domains. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Law (ICAIL-2001).* The Association for Computing Machinery Press, 12–21.

[12] Christof Benzmüller, Xavier Parent, and Leon van der Torre. 2020. Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. *Artificial Intelligence* 287 (2020), 103348.

[13] L. Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explanable legal prediction. *Artificial Intelligence and Law* 29 (2021), 213–238.

[14] Ilaria Canavotto and John Horty. 20xx. Reasoning with a hierarchy of open-textured predicates. (20xx). Unpublished manuscript.

[15] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First National Conference on Artificial Intelligence (AAAI-17).* 4831–4835.

[16] Jonathan Dancy. 1999. Can particularists learn the difference between right and wrong?. In *Proceedings of the Twentieth World Congress of Philosophy.* Philosophy Documentation Center, 59–72.

[17] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14.

[18] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John Dickerson, and Vincent Conitzer. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283 (2020), 1–14.

[19] Jean-Gabriel Ganascia. 2007. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology* 9 (2007), 39–47.

[20] David Gauthier. 1986. *Morals by Agreement.* Oxford University Press.

[21] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Williams. 2016. Embedding ethical principles in collective decision support systems. In *Proceedings of the Thirtieth National Conference on Artificial Intelligence (AAAI-16).* 4147–4151.

[22] John Horty. 2011. Rules and reasons in the theory of precedent. *Legal Theory* 17 (2011), 1–33.

[23] John Horty. 20xx. The Logic of Precedent: Constraint and Freedom in Common Law Reasoning. Forthcoming with Cambridge University Press.

[24] John Horty and Trevor Bench-Capon. 2012. A factor-based definition of precedential constraint. *Artificial Intelligence and Law* 20 (2012), 181–214.

[25] Edward Levi. 1949. *An Introduction to Legal Reasoning.* The University of Chicago Press.

[26] Sanjay Modgil and Henry Prakken. 2014. The ASPIC+ framework for structured argumentation: a tutorial. *Argument and Computation* 5 (2014).

[27] Luis Moniz Pereira and Ari Saptawijaya. 2016. *Programming Machine Ethics.* Springer.

[28] Henry Prakken. 2002. An exercise in formalising teleological case-based reasoning. *Artificial Intelligence and Law* 10 (2002), 113–133.

[29] Joseph Raz. 1975. *Practical Reason and Norms.* Hutchinson and Company. Second edition with new Postscript printed in 1990 by Princeton University Press, and reprinted by Oxford University Press in 2002; pagination refers to the Oxford edition.

[30] Joseph Raz. 1979. *The Authority of Law.* Oxford University Press.

[31] Raymond Reiter. 1980. A Logic for Default Reasoning. *Artificial Intelligence* 13 (1980), 81–132.

[32] Mark Riedl and Brent Harrison. 2016. Using Stories to Teach Human Values to Artificial Agents. In *Proceedings of the 2nd International Workshop on AI, Ethics and Society.*

[33] Edwina Rissland and Kevin Ashley. 1987. A case-based system for trade secrets law. In *Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL-87).* The Association for Computing Machinery Press, 60–66.

[34] W. D. Ross. 1930. *The Right and the Good.* Oxford University Press.

[35] Francesca Rossi. 2016. Moral preferences. (2016). Available at http://www.mpref-2016.preflib.org/wp-content/uploads/2016/06/paper-15.pdf.

[36] Alvin Roth, Tayfun Sonmez, and M. Utku Ünver. 2004. Kidney exchange. *Quarterly Journal of Economics* 119 (2004), 457–488.

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

[38] T. M. Scanlon. 1998. *What We Owe to Each Other.* Harvard University Press.

[39] Frederick Schauer. 1991. *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and Life.* Oxford University Press.

[40] Marek Sergot, Fariba Sadri, Robert Kowalski, Frank Kriwaczek, Peter Hammond, and H. Therese Cory. 1986. The British Nationality Act as a logic program. *Communications of the Association for Computing Machinery* 29 (1986), 370–386.

[41] Walter Sinnot-Armstrong and Joshua August Skorburg. 2021. How AI can aid bioethics. *Journal of Practical Ethics* 9 (2021).

[42] Joshua Skorburg, Walter Sinnott-Armstrong, and Vincent Conitzer. 2020. AI methods in bioethics. *AJOB Empirical Bioethics* 11 (2020), 37–39.

[43] Daniel Star. 2018. *The Oxford Handbook of Reasons and Normativity.* Oxford University Press.

[44] Wijnand Van Woerkom, Davide Grossi, Henry Prakken, and Bart Verheij. 2022. Landmarks in Case-based Reasoning: From Theory to Data. In *Proceedings of the First International Confefence on Hybrid Human-Machine Intelligence.* IOS Press.