The Jiminy Advisor:
Moral Agreements Among Stakeholders Based on
Norms and Argumentation

Beishui Liao baiseliao@zju.edu.cn
Marija Slavkovik marija.slavkovik@uib.no
Leendert van der Torre leon.vandertorre@uni.lu

January 14, 2022

Abstract
An autonomous system is constructed by a manufacturer, operates in a society subject to norms and laws, and is interacting with end users. All of these actors are stakeholders affected by the behavior of the autonomous system. We address the challenge of how the ethical views of such stakeholders can be integrated in the behavior of the autonomous system. We propose an ethical recommendation component, which we call Jiminy, that uses techniques from normative systems and formal argumentation to reach moral agreements among stakeholders. Jiminy represents the ethical views of each stakeholder by using normative systems, and has three ways of resolving moral dilemmas involving the opinions of the stakeholders. First, Jiminy considers how the arguments of the stakeholders relate to one another, which may already resolve the dilemma. Secondly, Jiminy combines the normative systems of the stakeholders such that the combined expertise of the stakeholders may resolve the dilemma. Thirdly, and only if these two other methods have failed, Jiminy uses context-sensitive rules to decide which of the stakeholders take preference. At the abstract level, these three methods are characterized by the addition of arguments, the addition of attacks among arguments, and the removal of attacks among arguments. We show how Jiminy can be used not only for ethical reasoning and collaborative decision making, but also for providing explanations about ethical behavior.

1 Introduction

Artificial autonomous systems depend on human intervention to distinguish moral from immoral behavior. Implicit ethical agents (Moor 2006) are ethically constrained
from engaging in immoral behavior via rules set by human designers. Explicit ethical agents (Moor [2006]; Dyrkolbotn et al. [2018]) or agents with functional morality (Wallach and Allen [2008], Chapter 2) are either able to make moral judgments themselves or are given guidelines or examples regarding what is good and bad. In either case, the question arises: who decides which and whose morality the artificial autonomous systems ultimately upholds?

It is immediately apparent that the persons and institutions who are affected by the moral behavior of an autonomous system should be given the opportunity to indicate their moral preferences as input into the behavior of that autonomous system (Baum [2020]). There are, however, numerous stakeholders satisfying the above definition of concerned entities (Baum [2020]). For example, governmental regulators can determine which behavior is legal for the state in which the autonomous system is deployed. The manufacturers, shareholders, designers and developers involved in building the autonomous systems would be concerned not only with issues of liability, but also issues of representation—an autonomous system should uphold the image and values of its maker. People interacting directly with the autonomous system should have a choice on certain aspects of its moral behavior, whether they are owners, users, or just share an environment with the system. It is easy to argue that it is wrong to select any of these stakeholders over others as being able to exclusively define what constitutes moral behavior for an autonomous system.

Legal systems recognize only humans and corporations as persons and moral agents. The underlying assumption that everyone is human allows a great deal of flexibility when it comes to specifying and enforcing desirable behavior—not all desirable behaviors are specified and not all violations of law are meticulously prosecuted. Companies can build a system that reports on every violation of law committed by a user, but who would then want to buy such a totalitarian “surveillance” device?

People can have multiple roles when interacting with an autonomous system, and they can have different role-based moral preferences for the system. As pedestrians, they would prefer utilitarian cars that elect to run into a wall and kill its one passenger rather than kill several pedestrians, yet at the same time would surely prefer not to buy such a car (Shariff et al. [2017]). Even if we somehow determine that the role of a pedestrian is more important than the role of a passenger, who would wish to buy a car that might kill one’s own children while driving them to school?

We propose here that all the stakeholders’ moral instructions should be included when deciding the moral behavior of an autonomous system. The problem then immediately becomes: how should an autonomous system dynamically combine the ethical input of various stakeholders?

In this paper, the terms “moral” and “ethical” are used interchangeably. Let us imagine that each of the “morality” stakeholders are represented by an “avatar” in
an artificial autonomous system. We refer to this artificial autonomous system as simply an “agent”. The “avatars” are the “moral council” of the agent, acting like Jiminy Cricket in the story of Pinocchio. First, the stakeholders can indicate which situations—decision contexts, actions or outcomes—are ethically sensitive. An agent makes such a decision by choosing from a set of available actions. If none of the stakeholders regard the situation as ethically sensitive in any way, then the agent can use its regular reasoning methods to select what to do. However, in an ethically sensitive situation, Jiminy would be employed to produce a moral recommendation to the agent.

The first challenge in building the “moral council” is that the stakeholders may not be following the same ethical reasoning theory or any ethical theory at all. It is not sufficient that each stakeholders agent chimes in with a “yes” or “no” when the question of the morality of an action or of an action’s outcome is presented.

The second challenge is that dilemmas and conflicts will arise when the inputs of the stakeholders are applied to a decision-making problem (Robinson, 2021; Horty, 1994). We do not want to evaluate the morality of an action by majority rule. Neither do we want to always put legal considerations above the image of the manufacturer, or give higher consideration to the personal input of the end user than the guidelines of regulatory bodies. Instead, we wish to have an engine that is able to take inputs from the different stakeholders and bring them into agreement.

The third challenge is explainability. Since the stakeholders are not necessarily aware of the input of other stakeholders, the decisions that the system ends up making need to be explained. That means that whatever solutions are used must be such that the artificial moral agent is able to explain its choices (Anderson and Leigh Anderson, 2014) or that the choices should be formally verifiable (Bremner et al., 2019).

We propose that normative systems (Chopra et al., 2018) and formal argumentation (Baroni et al., 2018a) can be used to implement a “moral council” for an artificial moral agent. With this approach, we can abstract away from how a particular stakeholder has reached a particular decision concerning the morality of an action. We model each stakeholder as a normative system that is then exploited as a source of arguments. An argument can be a statement regarding whether or not an action is moral, or it can be a reason why a particular action should be considered to be moral or immoral. Abstract argumentation allows us to build a system of attacking and supporting arguments that can be analyzed to determine which statements are supported and which are refuted in the system at a given time. Such a system can also generate explanations of decisions using dialogue techniques.

The main contributions of this paper are the following:

1. Within the field of machine ethics, this is the first computational model that
combines the ethical theories of multiple stakeholders in ethical decision making;

2. Within the field of structured argumentation, this is the first model that resolves moral dilemmas arising from multiple normative systems.

The paper is structured as follows. In Section 2 we introduce the Jiminy moral advisor component, in Section 3 we discuss how to represent normative systems, and in Section 4 we show different ways to use argumentation to come to moral agreements among stakeholders.

2 The Jiminy moral advisor component

We first consider the problem of how a multiple-stakeholder ethical advisory component can be designed and integrated into an artificial autonomous system. We call this multiple-stakeholder ethical advisory component a Jiminy advisor and refer to an artificial autonomous system simply as an agent.

The first problem we face is the problem of building the “avatars”, one for each stakeholder, and which are the sources of “insight” regarding what Jiminy should advise in a given ethically sensitive situation. Clearly, the stakeholders cannot be available in real time to give feedback to each instance of a Jiminy integrated within an agent. Rather, we need to obtain from them domain-specific information about what they consider to be ethically sensitive situations and their recommendations as to what should be done when such situations arise. We propose using normative systems (Chopra et al., 2018) to model the stakeholders. A normative system describes how to evaluated actions in a system of agents and how to guide the behavior of those agents (Alchourron, 1991). A norm is a formal description of desirable behavior, desirable action or the desirable outcome of an action. Furthermore, normative systems can also be seen as rule-based systems in which norms can be provided with reasons for supporting their enforcement. Besides presenting norms that stipulate how to avoid immoral behavior, stakeholders also contribute standpoints or claims in order to characterize and help identify ethically sensitive situations. Every stakeholder is modeled with their own normative system in the Jiminy advisor.

The advantage of using normative systems is that it allows us to abstract away from the particular moral theory that a stakeholder upholds. The immediate disadvantage of this approach is that the scope of ethically sensitive situations that the agent can handle by design are limited since these would need to be predicted in advance by the stakeholders. However, this is not unusual for systems that regulate behavior—even people sometimes find out that what they have done is immoral
after the fact. And even the law is not written to predict all future possible sources of danger to society. The law is subject to interpretation by legal practitioners, and is amended as necessary when a new threat is recognized. In the same way, a normative system can be amended if a new ethically sensitive situation arises.

It is clear that given the same ethically sensitive situation, different stakeholders would have different recommendations on what is the moral thing to do, supported by different reasons. As an example, imagine a smart house that detects the smoke of marijuana in a teenager’s room (Bjørgen et al., 2018). If the house is located in a state that criminalizes the use of marijuana, the stakeholder representing the state would argue that a crime has been committed that needs to be reported to the police. The stakeholder representing the house owners and parents of the teenager would argue that the misbehavior of the teenager is not something that should involve the law. Assume that the smart house is not a private home, but a hospital. Smoking marijuana might be allowed in certain parts for certain patients for medical reasons, but not in others. All of these “arguments” from the stakeholders are now available to the Jiminy advisor as normative systems.

Normative systems are built in different ways, depending on the size and complexity of the system. Relatively simple systems are built using regulative norms only, that directly relate a context with an obligation. If the system becomes more complicated and more contexts need to be distinguished, constitutive norms are used to define intermediate concepts. For example, there may be several constitutive norms that define what it means to get married, and then some regulative norms define the rights and duties that come with marriage. Intermediate concepts can be used to encode a decision tree between context and obligation. Finally, the general case is often distinguished from exceptional cases, and the latter are described using permissive norms. The role of the different kinds of norms is described in more detail in the following section.

We still need a way to bring together all pertinent norms and extract the moral recommendation that is best supported by the available arguments. Furthermore, the Jiminy advisor should produce an explanation as to why one particular action was recommended over another.

The relation between conditional norms on the one hand, and the obligations, permissions and institutional facts that follow from it, is known as detachment. In monotonic systems that cannot deal with conflicts, detachment corresponds to modus ponens in classical logic. With also permissive and constitutive norms, there are some choices to be made, as we explain in the next section. In particular, whether rules of different kinds can be applied after each other, and whether we can reason by cases. But once we consider nonmonotonic systems with some kind of built-in conflict resolution mechanism, the choice increases.

We propose using formal argumentation (Baroni et al., 2018a) to reach moral
agreements from stakeholders’ inputs. Formal argumentation is typically based on logical arguments constructed from prioritized rules. The first applications of formal argumentation in the area of normative multi-agent systems concerned the resolution of conflicting norms and norm compliance. Several frameworks have been proposed for normative and legal argumentation (Bench-Capon et al., 2010), but no comprehensive formal model of normative reasoning from arguments has yet been proposed.

Intuitively, an argumentation system consists of a set of arguments and a defeat relation over these arguments. Arguments can be constructed from an underlying knowledge base represented by a logical language, while the defeat relation can be defined in terms of the inconsistency of the underlying knowledge. Typically, an argumentation system is represented as a directed graph in which the nodes are arguments and there is an edge from node $A$ to node $B$ if argument $A$ attacks argument $B$ (see Figure 1). To find agreements, we consider that all possible arguments in the graph can be either admissible or inadmissible. An argument can only be admissible if all its attackers are inadmissible, or it has no attackers. An extension of an argumentation graph is any set of arguments that can be accepted together. For example, for the argumentation graph on Figure 1, there is only one possible extension, namely $\{A, C\}$. If the arguments contain moral recommendations (and supporting reasons), then the extension would contain an “agreed” unopposed moral recommendation from Jiminy to the moral agent.

The advantage of using the argumentation approach to reaching agreements is that it is fairly straightforward to generate explanations for agreements, as shown in Section 5. The disadvantage of the approach is that it is not always possible to arrive at only one possible agreement as to what is the most moral course of action, and two or more options can be equally justified as constituting an agreement. However, the disadvantage of possible ties is shared with other agreement reaching methods like social choice (Brandt et al., 2016), and is balanced against the benefit of easy access to explanations.

Having settled on how to represent the stakeholders and how to reach agreements on moral recommendations, we can illustrate the reasoning cycle of the Jiminy moral advisor in Figure 2. When Jiminy is triggered, it means that the agent is in an ethically sensitive situation. If the ethically sensitive situation can clearly be resolved, this is done directly. For example, let the agent be a smart house that manages a hospital, and non-marijuana smoke is detected. The agent has two choices: sound the fire alarm, or alert the nurses’ station. Both choices are passed to Jiminy as
available options. Assume further that all the stakeholders have recommended that in this situation the alarm should start sounding. This is what Jiminy returns as its moral recommendation to the agent: sound the alarm.

Consider, however, that where the ethically sensitive situation cannot be resolved from the options of what the agent can do, it means that none are considered moral by all the stakeholders’ normative systems. Now Jiminy uses the normative systems representing the stakeholders, together with any additional information from the agent’s knowledge base and its sensory input, to build the appropriate argumentation graph. Using this graph, Jiminy calculates the extensions from which it extracts the moral recommendations as well as the justifications for these recommendations, and returns both to the agent.

We provide details on the normative systems approach we use in Jiminy in Section 3 and discuss the argumentation reasoning approach in Section 4.

3 Representing normative systems, moral dilemmas and normative conflicts

We first distinguish logic-based from table-based normative systems, and we discuss two alternative logical languages suitable for logic-based normative systems. Then we discuss the representation of regulative and constitutive norms, and the related representation of moral dilemmas and normative conflicts. Finally, we discuss the
resolution of moral dilemmas and normative conflicts, in particular for systems with multiple stakeholders.

### 3.1 Table-based versus logic-based representations of normative systems

In its basic form, a normative system is a table expressing a relation between situations and deontic decisions. A prototypical example is a judge who decides on a verdict based on evidence. In the case of ethical agents, a knowledge engineer can present each stakeholder with a table like Table 1 to complete. Each row in the table can be seen as a simple norm, e.g., “In situation 1, you should alert the household”. In the table-based representation, a normative system is thus a set of such simple norms.

| Situation description | Recommended decision |
|-----------------------|----------------------|
| Alert authorities     | Alert household      |
| Situation 1           | x                    |
| Situation 2           | x                    |
| ...                   |                      |
| Situation m           | x                    |

Table 1: A very simple normative system in table form

A normative system table can also be elicited via a web interface presenting scenarios of moral dilemmas and having the stakeholder select from alternative options. For example, in the moral machine experiment (Awad et al., 2018), the user is presented with a number of scenarios, and for each scenario, has to choose between two alternatives. Though this table-based method is very simple and thus easy for the stakeholder to understand and use, it is not very efficient from the perspective of knowledge engineering because the number of situations is fixed beforehand and typically has to remain small.

A more advanced representation, often attributed to Ross [1957], is to represent a normative system by using two tables (see, for example, Table 2). The first table, Table 2a, relates situations or contexts to a set of features or factors, and the second table, Table 2a, relates these features to deontic decisions. There are now two kinds of norms. **Constitutive Norms** relate situations to features, e.g., “Situation 1 counts as Feature 1 and Feature 5”, while **Regulative Norms** relate features to recommendations, e.g., “if Feature 1 and Feature 5 apply, then Alternative 1 should be chosen”. A normative system is a set of constitutive and regulative norms. For a recent discussion of this representational technique, see the work of Grossi and Jones [2013]
The features may refer to more abstract legal terms such as blasphemy, privacy, contract, or ownership. In the ethical agent architecture, the ontology of these features may be shared by all stakeholders. Depending on the application domain, the features are called intermediate or institutional facts in order to distinguish them from the propositions used to describe the situations, which are called brute facts. The same structure that is used for constitutive and regulative norms has also been used for practical or goal-based reasoning. In that case, the intermediate facts may refer to goals or desires ([Broersen et al., 2001]).

Notwithstanding the distinction between constitutive and regulative norms, often permissive norms are regarded as distinct in a normative system. They have the same structure as regulative norms, but are used to describe exceptions. For example, there can be a general norm prescribing client confidentiality, but confidentiality can be broken when clients represent a threat to themselves or others (see, for example, Table 3).

The logic-based representation of normative systems further refines the table-based representation in order to increase representational efficiency[^1]. For example, the combination of features in Tables 2 and 3 corresponds to the logical conjunction

[^1]: Algebraic formalisms have been used widely for the same reason, e.g., by Lindahl and Odelstad 2013.
of literals. If the table is represented by logical formulas instead of a list, other connectives can also be used, such as logical disjunctions. Several rows in the table can then be represented with a single formula. For example, if Situation \( i \) or Situation \( j \) then Feature \( k \) and \( l \). Alchourrón and Bulygin [1981] developed their logic-based representation of normative systems inspired by the Tarskian theory of deductive systems, i.e., mathematical proof theories in deductive logic. The logic-based representation is often based on a nonmonotonic logic because the norms can be subject to exceptions due to, for example, permissive norms. In this paper, we use formal argumentation for the nonmonotonicity inherent in normative reasoning.

We now provide Definition 1 of a normative system. It formalizes regulative, constitutive and permissive norms. We assume that the agents have the same features. In this paper, we assume a shared unique language \( \mathcal{L} \) based on a shared set of propositional atoms \( \mathcal{P} \). Definition 1 represents a relatively abstract theory of normative systems, and we believe that it is precisely this generality that makes normative systems suitable for the Jiminy architecture. Instead of using negation, we adopt the more general concept of a contrariness function. This contrariness function is not necessarily symmetric and is therefore more general than standard negation. It is popular in formal argumentation and is a generalization of weak negation in logic programming. For the general theory of generalized contradiction in formal argumentation, see the work of Baroni et al. [2018b].

**Definition 1 (Normative system).** Given a set of stakeholders containing a stakeholder \( s \), the normative system of \( s \) is a tuple \( \mathcal{NS} = (\mathcal{L}, \neg, \mathcal{R}) \) where,

- \( \mathcal{L} \) is a logical language containing for all stakeholders \( s_1 \) and \( s_2 \) an atomic formula of the form \( s_1 > s_2 \) to indicate that stakeholder \( s_1 \) is superior to stakeholder \( s_2 \).

- \( \neg : \mathcal{L} \mapsto 2^\mathcal{L} \) is a so-called contrariness function, such that \( s_2 > s_1 \in \overline{s_1 > s_2} \) for all stakeholders \( s_1 \) and \( s_2 \).

- \( \mathcal{R} \) is a set of norms whose elements are of the form \( \phi_1, \ldots, \phi_n \Rightarrow^\tau \phi \), where \( \phi_1, \ldots, \phi_n, \phi \in \mathcal{L} \) and \( \tau \in \{r, c, p\} \). \( \mathcal{R}^r, \mathcal{R}^c \) and \( \mathcal{R}^p \) contain the norms with the corresponding superscript and are called regulative norms, constitutive norms and permissive norms respectively. They are indicated by \( r \), \( c \) and \( p \) respectively.

We write \(- (\psi, \phi) \) iff \( \psi \in \overline{\phi} \) and \( \phi \in \overline{\psi} \). If \( \neg \) is part of the logical language, and \( p \) is a propositional atom, then we require \( \{p\} \in \overline{\neg} \) and \( \{\neg p\} \in \overline{p} \), i.e., \(- (\neg p, p) \).
3.2 The choice of logical language in logic-based normative systems

We can adopt a classical propositional, first-order or a modal language. A modal language can contain modal operators for obligation, permission and prohibition. For example, Standard Deontic Logic (SDL) is a normal propositional modal logic of type KD, which means that it extends the propositional tautologies with the axioms $K: O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$ and $D: \neg(Op \land O\neg p)$, and it is closed under the inference rules modus ponens, $p, p \rightarrow q/q$, and necessitation, $p/Op$. Prohibition and permission are defined by $Fp = O:\neg p$ and $Pp = :Op$. SDL is an unusually simple and elegant theory. In this section, we discuss the pros and cons of these two options.

It may be observed that some authors in deontic logic use the concept of norm and conditional obligation interchangeably. However, the distinction between norms and obligations was articulated by Makinson 1999 and further developed formally in input/output logic (Makinson and van der Torre 2000). To detach an obligation from a norm, there must be a context, and the norm must be conditional. Thus norms as defined in Definition 1 are just particular kinds of rules, and one may view a normative system simply as a set of rules.

Since modal logic integrates classical logic just like first-order logic integrates propositional logic, it may be argued that we can express more when we use a modal language as the base language. For example, there are examples where permissions give rise to new obligations and permissions, and this can be expressed only when we adopt a modal language.

However, examples of where we need the expressive power of a modal language are rare. For most practical purposes, a classical language may be sufficient. As Makinson explains, the absence of explicit modal operators in normative systems may be seen as a limitation, but it also facilitates formal analysis. Makinson attributes the “liberating effect” of no longer having to explicitly represent the modal operator to Alchourrón and Bulygin 1981:

An unconditional normative code is defined to be a pair $N = (A, B)$ where $A, B$ are sets of purely boolean formulae. Intuitively they represent, respectively, the states of affairs that the code explicitly requires to come into effect, and those that it explicitly permits to do so.

---

2Not surprisingly for such a highly simplified theory, there are many features of actual normative reasoning that SDL does not capture. For example, the Handbook of Deontic Logic and Normative Systems (Gabbay et al. 2013) explains in detail the so-called ‘paradoxes of deontic logic’, which are usually dismissed as consequences of the simplifications of SDL. For example, Ross’s paradox (Ross 1941) where the counterintuitive derivation of “you ought to mail or burn the letter” from “you ought to mail the letter”, is typically viewed as a side effect of the interpretation of ‘or’ in natural language.
There is thus a small, but immensely significant step compared to the sketch of Stenius [1963]. Alchourrón and Bulygin appear to have been the first to realise the liberating effect of taking the set of promulgations of a normative code to be made up of purely boolean formulae. At the same time, they consider explicit permissions along with promulgations (Makinson [1999], p. 32-33).

In this paper, we use classical logic as the base logic, but all our definitions involve a modal logic as well.

### 3.3 Representing constitutive and regulative norms

In this section, we provide some guidelines on how to represent constitutive and regulative norms, and we demonstrate the representation of such norms with a running example. Normative systems have been used in many disciplines. Consequently, besides the relatively abstract theory in Definition 1 which can be used across disciplines, there are also more detailed theories that have been developed to be used in specific disciplines because they have been adapted to the specific concerns of those disciplines.

Constitutive norms are rules that create the possibility of undertaking an activity, or rules that define an activity. For example, according to Searle [1969], the activity of playing chess is constituted by action in accordance with the rules of the game. The institutions of marriage, money, and promising are like the institutions of baseball and chess in the sense that they are all systems of constitutive rules or conventions. As another example, a signature may count as a legal contract, and a legal contract may define both a permission to use a resource and an obligation to pay. Searle points out that, unlike regulative norms, constitutive rules do not regulate actions but define new forms of behavior. Constitutive norms link brute facts (like the signature of a contract) to institutional facts (a legal contract), and are usually represented as counts-as conditionals: $X$ counts as $Y$ in context $C$. Searle’s analysis insists on the contextual nature of constitutive norms: a signature counts-as a legal contract when written on a paper stating the terms of such a contract. If, however, I write my signature on a white sheet, that does not constitute a legal contract. Constitutive norms have been identified as the key mechanism of normative reasoning in dynamic and uncertain environments, for example to achieve agent communication, electronic contracting, and handle the dynamics of organizations (see, for example, the work of Boella and van der Torre [2006]).

Regulative norms, including permissive norms, indicate what is obligatory or permitted. In formal deontic logic, permissions are studied less frequently than obligations. For a long time, it was naively assumed that a permission could simply
be taken as a dual of obligation, just like possibility is the dual of necessity in modal logic. However, Bulygin [1986] observed that an authoritative kind of permission must be used in the context of multiple authorities and dynamic normative systems: if a higher authority permits you to do something, a lower authority can no longer prohibit it. Deontic logic has been concerned mainly with regulative norms, but the logic of constitutive norms [Grossi and Jones 2013] is a subject of study of its own.

**Example 3.1.** As a running example, consider a simple morally sensitive situation where the following five norms are used in moral decision making:

1. If you have manufactured a device, the behavior of that device should comply with the law. (Stakeholder $M$)
2. You are permitted to collect user data without explicit permission from the user. (Stakeholder $M$)
3. Devices that collect user data should protect the privacy of their users. (Stakeholder $H$)
4. Devices that contain information about a future event that grossly endangers society should report that information to the authorities. (Stakeholder $H$)
5. Choosing to do business in Norway counts as requiring compliance with the General Data Protection Regulation 2016/679 (GDPR). (Stakeholder $L$)

In this morally sensitive situation, the context is:

- (w₁) the manufacturer makes the smart speaker,
- (w₂) the smart speaker collects user data,
- (w₃) the information collected indicate a potential critical danger to society (for example a situation in which many lives are lost),
- (w₄) the manufacturer does business in Norway.

There are three conflicting moral options:

- (d₁) comply with the law,
- (d₂) protect the privacy of users,
- (d₃) report information concerning a potential critical danger to society.
3.4 Moral dilemmas involving conflicting obligations

In general, moral dilemmas are situations where it is no longer possible to satisfy all norms, i.e., at least one norm must be violated. The representation of violations is built on the distinction between “is” and “ought.” David Hume introduced the so-called is-ought problem, which roughly means that there is a fundamental difference between positive statements and prescriptive or normative statements. The is-ought problem can be considered in two directions. First, what is the case cannot be the basis for what ought to be the case. Second, what ought to be the case cannot be the basis for what is the case. This is related to the fallacy of wishful thinking: an agent may want to win the lottery, but from that desire he should not deduce that he will win the lottery. Likewise, an agent should not, in a kind of deontic wishful thinking, deduce from the mere fact that she is obliged to review a paper that she will actually do it. We call agents realistic if they do not make such fallacious inferences. The fundamental distinction between “is” and “ought” is the main reason why deontic logic is normally formalized as a branch of modal logic. It distinguishes brute facts like \( p \) from deontic facts like obligations \( O p \) and permissions \( P p \), and it represents violations with mixed formulas like \( p \land O \neg p \).

As a first approximation, one may be tempted to define moral dilemmas as two conflicting norms. For example, if there are two norms, one prescribing alerting the police and the other prescribing not alerting the police, and the condition for the activation of both norms is part of the current context, then we might be tempted to deduce that there is a moral dilemma. However, the problem with this definition of a moral dilemma in terms of a normative system is that the norms are defeasible. So even if there is a norm prescribing alerting the police and another norm prescribing not alerting the police, there may also be a permission implying an exception to one of these norms. If there is such a permission, there is no longer a moral dilemma.

For this reason, moral dilemmas in normative reasoning are usually not defined in terms of the normative system, but in terms of the conclusions of the norms that are detached from the normative system. There are two approaches to detachment, depending on the choice of logical language used for the logic-based normative systems described in Section 3.2. In the first approach, where the logical language is a modal logic containing at least a modal operator \( O \) for obligation, a moral dilemma is represented by an unresolved conflict between two incompatible obligations, e.g., \( O p \land O \neg p \). With the deontic axiom \( \neg (O p \land O \neg p) \), so-called Standard Deontic Logic makes deontic dilemmas inconsistent, but many alternative logics allow consistent representation of such dilemmas and thus reject this axiom. In the second approach, where the logical language does not contain a modal operator, moral dilemmas are usually represented by detachment of so-called extensions. An extension is a consistent set of formulas pertaining to the logical language. Whereas
in most logics, we can derive only a single set of conclusions from a set of premises, in normative reasoning there may be several such sets. If there is more than one extension, then this indicates some kind of conflict. In the formal argumentation adopted in this paper, moral dilemmas are also represented by the existence of multiple extensions.

3.5 Normative conflicts among conflicting institutional facts

There is one additional challenge when defining moral dilemmas due to the existence of constitutive norms in terms of multiple extensions. We can derive conflicting institutional facts from a normative system, and this will also lead to multiple extensions. To distinguish this situation from conflicting obligations, we call such conflicting institutional facts a normative conflict.

Since the concepts of moral dilemma and normative conflict are defined in terms of the detachment procedure, this section gives only an informal characterization of the distinction between moral dilemma and normative conflict. We will formally define them in the next section when we have defined the detachment procedure based on formal argumentation.

**Moral dilemma** Multiple extensions due to conflicting obligations. For example, one stakeholder believes that we should alert the police while another stakeholder believes that we should alert the parents.

**Normative conflict** Multiple extensions due to conflicting institutional facts. We call this normative conflict but not normative or moral dilemma. For example, one stakeholder may believe that a certain situation counts as blasphemy, whereas another agent believes that the same situation does not count as blasphemy. This is a disagreement about the nature of the situation, not explicitly about the actions to be taken.

Normative conflicts may lead to moral dilemmas. For example, if one stakeholder believes that there has been blasphemy while another stakeholder does not (a normative conflict), the first stakeholder may deduce that we should alert the police while the second stakeholder may not (a moral dilemma).

Reasoning about moral dilemmas and normative conflicts should not be confused with contrary-to-duty reasoning that concerns the representation of consequences of violations such as sanctions and reparations. A contrary-to-duty obligation expresses what one should do when obligations have been violated. In other words, contrary-to-duty obligations are triggered by conflicts between what is the case and what ought to be the case, and they may be seen as a way of resolving this conflict, if only partially. Of course, it is better to review a paper than not doing
the review and being sanctioned for that. Many deontic logic paradoxes contain contrary-to-duty obligations, such as the gentle murderer paradox: a person should not kill, but, if he kills, he should do it gently. Such scenarios should be represented in a consistent way, but in many deontic logics, such formalisations are inconsistent or have counterintuitive consequences.

3.6 Moral agreement: resolving moral dilemmas and normative conflicts

In hierarchical normative systems, conflicts among norms can be resolved by reference to the hierarchy, which can be based on the authority that promulgated the norm, but which can also refer to other information such as the time of the promulgation, or the specificity of the norm. In this paper, we do not hardcode a global ordering on stakeholders, purposes, or values, as no agreement may exist on such an ordering. This is comparable to the status of autonomous countries in international law, where it is assumed that there is no order among the countries.

Each agent can have normative conflicts and/or moral dilemmas. In the case of multiple stakeholders, there can be four levels of normative conflicts and/or moral dilemmas:

1. **Stakeholder dilemma/conflict** Stakeholders accept distinct arguments when considered in isolation from other stakeholders;

2. **Large dilemma/conflict** When all arguments are combined in a large framework, there are multiple extensions;

3. **Big dilemma/conflict** When all normative systems are combined to generate a big framework, there are multiple extensions;

4. **Jiminy dilemma/conflict** When considering one of the above (individual frameworks, a large framework or the big framework) together with stakeholder selection norms, there are still multiple extensions.

We say that a moral dilemma or normative conflict is resolved when at some levels there are multiple extensions, but at a higher level there is only one. So if, due to the stakeholder selection norms, there is only one extension at level 4, then we say that the stakeholder selection norms resolve the moral dilemma or normative conflict.

If some of the stakeholders find an event immoral, and others do not, then two kinds of discussions can be triggered. The first kind of discussion aims to question the moral judgment of another stakeholder. The moral judgment of stakeholders is typically based on assumptions, judgments and goals. For example, the moral decision to recommend calling the police may be based on the assumption that the
relevant persons are adults, the judgment that the discussion counts as blasphemy, and the goal of reporting blasphemy. Each of these elements can be questioned. For example, a stakeholder can claim that the assumption does not hold because the voices of children are detected, or that the discussion does not count as blasphemy, or that the goal to report blasphemy does not exist in the country where the discussion is held, or that there may be a more important goal of protecting the privacy of the household.

The second kind of discussion that can be triggered is a conflict resolution discussion. In a conflict resolution discussion, special norms can be used to decide which normative system is applicable to a particular situation. For example, there may be a norm that states that in the case of a life-threatening situation, the normative system of the law overrides the normative systems of other stakeholders. Such norms may be particular fragments of the legal code in international private law, for example.

We introduce a special normative system called \( J \) for Jiminy containing specific representations explaining which normative system is in use. The norms in this conflict resolution mechanism contain only strict constitutive norms.

The complexity of a conflict resolution argument is that the features that decide which normative system is applicable, like the existence of a life-threatening situation, may themselves be subject to debate. So one stakeholder may argue that a particular situation is life-threatening while another stakeholder may argue that it is not. In such cases, again a conflict resolution argument can be triggered, in this case not to resolve the ethical dilemma, but to agree on a collective judgment.

### 3.7 Detachment

Since there are no priorities associated with the norms, the detachment procedure is relatively straightforward. Given a context, we can apply constitutive norms iteratively, and then we can apply the regulative norms. The main choices to be made are as follows:

1. Do we allow reasoning by cases? For example, when we say that it is forbidden to use a radio in the park and it is forbidden to use a radio in the classroom, and we know that we are either in the park or in the classroom, do we detach that it is forbidden to use a radio? The drawback of reasoning by cases is that it complicates the inference relation and it increases the complexity, and therefore we do not adopt it.

2. Do we allow iterated detachment of obligations, known as deontic detachment? It is well known from deontic paradoxes like Chisholm’s paradox or Forrester’s
paradox that deontic detachment is problematic, and deontic logics that deal with it are computationally more demanding. Therefore we do not adopt it.

3. Do we allow the use of constitutive norms in the scope of obligations and permissions?

Based on the above discussion, we end up with the following definition of detachment.

**Definition 2.** Let $K$ be a set of formulas representing the context, let $N^c$ be a set of constitutive norms, $N^r$ be a set of regulative norms and $N^p$ be a set of permissive norms. Moreover, let $Cn$ be the consequence relation of the base logic $L$ containing at least the partial order properties of $\succ$, i.e. transitivity and asymmetry. For a set of norms $N$, we define a one-step application of the norms $N$ to $K$, written as $N^1(K)$, as follows:

$$N(K) = \{ x \mid a \Rightarrow x \in N, a \in Cn(K) \}$$

Based on this one-step detachment, we have the following:

- The institutional facts are the formulas that can be detached iteratively from $K$ and $N^c$: $I_0(N^c, K) = K$, $I_{i+1}(N^c, K) = I_i(N^c, K) \cup N^c(I_i(N^c, K))$, $I(N^c, K) = \cup I_i(N^c, K)$;

- The obligations are the formulas that can be detached from $I$ and $N^r$: $O(N^c, N^r, K) = N^r(I(N^c, K))$;

- The permissions are the formulas that can be detached from $I$ and $N^p$: $P(N^c, N^p, K) = N^p(I(N^c, K))$.

If we consider only the norms of an agent $\alpha$ of $N^c$, then we write $N^c_{\alpha}$ etc.

Roughly, there is a conflict when $I$ is inconsistent, there is a violation when $I$ is inconsistent with $O$, and there is a dilemma when $O$ is inconsistent or when there is a permission $p \in P$ such that $\{ p \} \cup O$ is inconsistent. We make this idea more precise using the notion of extension, in the sense that conflicts and dilemmas are represented by multiple extensions, and violations are filtered out of these extensions. Such extensions are represented both by sets of norms and by sets of formulas.

**Definition 3.** A norm extension of $(N^c, N^r, N^p)$ in context $K$ is a triple $(M^c, M^r, M^p)$ such that:

- $M^c$ is a maximal subset of $N^c$ such that $I(M^c, K)$ is consistent;

- $M^r$ and $M^p$ are maximal subsets of $N^r$ and $N^p$ such that for all $p \in P(M^c, M^p, K)$, we have that $O(M^c, M^r, K) \cup \{ p \}$ is consistent.
The norm extension \((M^c, M^r, M^p)\) corresponds to institutional facts \(I(M^c, K)\), obligations \(O(M^c, M^r, K)\) and permissions \(P(M^c, M^p, K)\).

In the following section, some of these conflicts and dilemmas are resolved using the contrariness function and an argumentation theory.

4 Argumentation-based moral agreements among stakeholders

In this section, we focus on how to check and resolve a moral dilemma by constructing, comparing and evaluating arguments at different levels in terms of a moral decision problem and a set of normative systems for representing the stakeholders. Firstly, we formalize the notions of morally sensitive situations, moral decision variables (or called deontic options) and moral decision problems as follows.

Definition 4 (Morally sensitive situation). Given the language \(L\) of the normative systems, a morally sensitive situation is defined as \(mss \in 2^L\). The set of all possible morally sensitive situations of an ethical agent is denoted as \(MSS\).

Morally sensitive situations of an ethical agent are given in advance.

Definition 5 (Moral decision variable). Given the language \(L\) of the normative systems, the set of all possible moral decision variables of the ethical agent is defined as \(DV \subseteq L\), representing the set of all possible deontic options the ethical agent might handle within the morally sensitive situations.

Given a morally sensitive situation and a set of moral options, a moral decision problem is about deciding which option should be selected, in terms of the norms of all stakeholders, as well as that of the Jimimy when needed.

Definition 6 (Moral decision problem). For each morally sensitive situation \(mss \in MSS\) and each subset of moral decision variables \(DV \subseteq DV\), \(DP = (mss, DV)\) is a moral decision problem.

Example 4.1. Consider Example [3.1]. The set \(\{w_1, w_2, w_3, w_4\} \subseteq L\) is a morally sensitive situation. These were: \((w_1)\) the manufacturer makes the smart speaker, \((w_2)\) the smart speaker is collecting user data, \((w_3)\) the information collected informs about a potential critical danger to society, and \((w_4)\) the manufacturer is doing business in Norway. The set \(DV = \{d_1, d_2, d_3\}\) is a set of moral decision variables. These were \((d_1)\) comply with the law, \((d_2)\) protect the privacy of users, and \((d_3)\) report the information concerning a potential critical danger to society. Then, \((\{w_1, w_2, w_3, w_4\}, \{d_1, d_2, d_3\})\) is a moral decision problem.
Given a set of normative systems and a moral decision problem, in the following, we formulate an argumentation-based approach for checking and resolving a moral dilemma. Arguments are constructed from an argumentation theory, which consists of a normative system as presented in Definition 1 and a knowledge base.

Definition 7 (Argumentation theory). Let \( S = \{s_1, \ldots, s_n, J\} \) be a set of stakeholders, where \( J \) stand for Jiminy.

An argumentation theory of a stakeholder \( s \in S \) is a tuple \( AT_s = (\text{NS}_s, \mathcal{K}) \), where \( \text{NS}_s = (\mathcal{L}, \overline{\mathcal{R}}, \mathcal{S}) \) is the normative system of \( s \), and \( \mathcal{K} \subseteq \mathcal{L} \) is the set of observations, called the context.

A combined argumentation theory of a set of normative systems of the stakeholders \( N \subseteq S \) is a tuple \( AT_N = (\text{NS}_N, \mathcal{K}) \). We have \( \text{NS}_N = (\mathcal{L}, \overline{\mathcal{R}}, \mathcal{S}) \) and \( \mathcal{R}_N = \bigcup_{s \in N} \mathcal{R}_s \).

Note that for the normative system of the Jiminy \( \text{NS}_J = (\mathcal{L}, \overline{\mathcal{R}}, \mathcal{S}) \), each norm in \( \mathcal{R}_J \) is of the form \( \psi_1, \ldots, \psi_k \Rightarrow s_1 > s_2 \), which denotes a context-sensitive rule used to decide which of the stakeholders takes preference. In the following, a set of stakeholders without \( J \) is denoted as \( S_0 \), i.e., \( S_0 = S \setminus \{J\} \).

Example 4.2. Let \( S = \{L, H, M, J\} \) be a set of stakeholders, where \( J \) is the Jiminy. Consider four argumentation and normative theories \( AT_s = (\text{NS}_s, \mathcal{K}), \text{NS}_s = (\mathcal{L}, \overline{\mathcal{R}}, \mathcal{S}), s \in S \), where:

- \( \mathcal{L} = \{w_1, w_2, w_3, w_4, a_1, a_2, d_1, d_2, d_3\} \cup \{s > s' \mid s, s' \in S\} \), in which \( w_1, w_2, w_3, w_4, d_1, d_2, d_3 \) are as in Example 3.1 while \( a_1 \) and \( a_2 \) denote:
  
  (a1) comply with the GDPR,
  (a2) collect user data without an explicit permission from the user;

- We have \( \bar{a}_1 = \{a_2\}, \bar{a}_2 = \{a_1\}, \bar{d}_1 = \{d_2, a_3\}, \bar{d}_2 = \{a_2, d_1\}, \bar{d}_3 = \{d_2\} \), from which we may infer that: \( a_1 = \neg a_2 \) and \( d_1 = \neg d_2 \), meaning that \( a_1 \) and \( a_2 \), respectively \( d_1 \) and \( d_2 \), are contradictory, while \( a_2 \in \bar{d}_2 \) and \( a_3 \in \bar{d}_1 \) mean that \( a_2 \) is a contrary of \( d_2 \) and \( a_3 \) is a contrary of \( d_1 \);

- \( s > s' = \{s' > s\}, s' > s = \{s > s'\} \) for all \( s, s' \in S \);

- \( \mathcal{R}_L = \{w_1 \Rightarrow_L d_1\}, \mathcal{R}_H = \{w_2 \Rightarrow_H d_2, w_3 \Rightarrow_H d_3\}, \mathcal{R}_M = \{w_4 \Rightarrow_M a_1, a_3 \Rightarrow_M a_2\}, \mathcal{R}_J = \{a_1 \Rightarrow L > H, a_2 \Rightarrow H > L\} \), where \( \mathcal{R}_L, \mathcal{R}_H \) and \( \mathcal{R}_L \) contain five norms described in Example 3.1 while \( a_1 \Rightarrow L > H \) and \( a_2 \Rightarrow H > L \) in \( \mathcal{R}_J \) denote dilemma resolving norms of the Jiminy, saying that under when \( a_1 \) holds, the stakeholder \( L \) is preferred to the stakeholder \( H \), and when \( a_2 \) holds, the stakeholder \( H \) is preferred to the stakeholder \( L \);
\( \mathcal{K} = \{w_1, w_2, w_3, w_4\} \) is the context.

In this paper, the notion of argument is defined in terms of the one presented by Pigozzi and van der Torre (\( \) ). Since we assume that all norms are defeasible, all arguments constructed from normative systems are defeasible. Moreover, the notion of norms used in the definition is corresponding to that of rules.

Informally speaking, an argument is a statement or a collection of statements that support(s) another statement. The former is called a premise (a set of premises), while the latter is called a conclusion. In a rule-based system, a conclusion of an argument can be derived from the premises by using a set of rules. Following Pigozzi and van der Torre (\( \) ), norms are used as rules to derive conclusions. So, arguments are constructed from given normative systems that are associated with one or more stakeholders. The set of arguments follows directly from Definition 2 of detachment in the normative system, as each argument is a derivation corresponding to a sequence of detachments. For an argument, we say that it is:

- a brute fact argument, if it contains facts from the knowledge base.
- an institutional fact argument, if it contains at least one instance of constitutive norms \( \mathcal{R}_c \) and no instances of regulative \( \mathcal{R}_r \) or permissive norms \( \mathcal{R}_p \). The set of institutional fact arguments is denoted \( \mathcal{A}_c \).
- an obligation argument, if it contains at least one instance of regulative norms \( \mathcal{R}_r \) and no instances of permissive norms \( \mathcal{R}_p \). The set of obligation arguments is denoted \( \mathcal{A}_r \).
- a permission argument if it contains at least one instance of permissive norms \( \mathcal{R}_p \). The set of permission arguments is denoted \( \mathcal{A}_p \).

**Definition 8 (Argument).** Let \( S \) be a set of stakeholders. Let \( AT_s = (NS_s, \mathcal{K}) \) be an argumentation theory of a stakeholder \( s \in S \) and \( AT_j = (NS_J, \mathcal{K}) \) be an argumentation theory of Jiminy. Let \( AT_N = (NS_N, \mathcal{K}) \) be a combined argumentation theory for \( N \subseteq S \). An argument \( A \) is:

1. a brute fact argument \( \phi \) if \( \phi \in \mathcal{K} \); \( \text{Conc}(A) = \phi \).
2. an institutional fact argument \( A_1, \ldots, A_n \Rightarrow^c_\mathcal{K} \phi \) if \( A_1, \ldots, A_n \) are brute or institutional fact arguments such that there exists a norm \( \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow^c_\mathcal{K} \phi \) in \( \mathcal{R}_s \); \( \text{Conc}(A) = \phi \).
3. an obligation argument \( A_1, \ldots, A_n \Rightarrow^r_\mathcal{K} \phi \) if \( A_1, \ldots, A_n \) are brute or institutional fact arguments such that there exists a norm \( \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow^r_\mathcal{K} \phi \) in \( \mathcal{R}_s \); \( \text{Conc}(A) = \phi \).
4. a permission argument \( A_1, \ldots, A_n \Rightarrow^p \phi \) if \( A_1, \ldots, A_n \) are brute or institutional fact arguments such that there exists a norm \( \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow^p \phi \) in \( R_s \); \( \text{Conc}(A) = \phi \).

5. a dilemma resolving argument \( A_1, \ldots, A_n \Rightarrow s_1 > s_2 \) if \( A_1, \ldots, A_n \) are brute or institutional fact arguments such that there exists a norm \( \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow s_1 > s_2 \) in \( R_J \); \( \text{Conc}(A) = s_1 > s_2 \).

Let \( R^t \subseteq R_s \cup \cdots \cup R_{sn} \), \( \tau \in \{ r, c, p \} \), be the set of institutional, regulative, and permissible norms, respectively. The conclusions of object level normative arguments are called institutional facts, obligations and permissions respectively. The sets of all arguments constructed from \( AT_s \) and \( AT_N \) are denoted as \( \text{Arg}(AT_s) \) and \( \text{Arg}(AT_N) \), respectively.

**Example 4.3.** We continue Example 4.2. We may construct the following arguments.

- \( A_{01} = w_1 \): the manufacturer makes the smart speaker.
- \( A_{02} = w_2 \): the smart speaker is collecting user data.
- \( A_{03} = w_3 \): the collected (user) information informs about a potential critical danger to society.
- \( A_{04} = w_4 \): the manufacturer is doing business in Norway.
- \( A_1 = (A_{01} \Rightarrow^r d_1) \): if the manufacturer makes the smart speaker, then they should comply with the law.
- \( A_2 = (A_{02} \Rightarrow^r d_2) \): if the smart speaker is collecting user data, then the privacy of the users should be protected.
- \( A_3 = (A_{03} \Rightarrow^r d_3) \): the collected (user) information informs about a potential critical danger to society, then this information should be reported.
- \( A_4 = (A_{04} \Rightarrow^r a_1) \): if the manufacturer is doing business in Norway, then they should comply with the GDPR.

Now we can form \( \text{Arg}(AT_L) = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_1 \} \), \( \text{Arg}(AT_H) = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_2, A_3 \} \), \( \text{Arg}(AT_M) = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_4 \} \).

We define some useful functions over arguments. Let \( A \) be an argument. The function \( \text{Prem}(A) \) returns the premises of the argument \( A \). The function \( \text{Conc}(A) \) returns the conclusion of the argument \( A \). The function \( \text{Sub}(A) \) returns the set of sub-arguments of \( A \). The function \( \text{Norms}(A) \) returns the set of norms used in
argument $A$. The function $\text{TopNorm}(A)$ returns the top norm used in $A$. Lastly, the function $\text{Stakeholder}(A)$ returns the set of stakeholders who supply the norms used in $A$.

**Definition 9 (Argument properties).** We use $\text{Arg}^{\text{Obl}}(AT_s) \subseteq \text{Arg}(AT_s)$ and $\text{Arg}^{\text{Obl}}(AT_N) \subseteq \text{Arg}(AT_N)$ to denote the sets of obligation arguments in $\text{Arg}(AT_s)$ and $\text{Arg}(AT_N)$, respectively. For an argument $A$ we define:

1. If $A$ is a **context argument** $\phi$ then:
   - $\text{Prem}(A) = \{ \phi \}$,
   - $\text{Conc}(A) = \phi$,
   - $\text{Sub}(A) = \{ \phi \}$, $\text{Norms}(A) = \emptyset$,
   - $\text{TopNorm}(A) = \text{undefined}$;
   - $\text{Stakeholder}(A) = \emptyset$.

2. If $A$ is a **normative argument** $A_1, \ldots, A_n \Rightarrow^\tau \phi$ then:
   - $\text{Prem}(A) = \text{Prem}(A_1) \cup \ldots \cup \text{Prem}(A_n)$,
   - $\text{Conc}(A) = \phi$,
   - $\text{Sub}(A) = \text{Sub}(A_1) \cup \ldots \cup \text{Sub}(A_n) \cup \{ A \}$,
   - $\text{Norms}(A) = \text{Norms}(A_1) \cup \ldots \text{Norms}(A_n) \cup \{ \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow^\tau \phi \}$,
   - $\text{TopNorm}(A) = \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow^\tau \phi$,
   - $\text{Stakeholder}(A) = \text{Stakeholder}(A_1) \cup \ldots \cup \text{Stakeholder}(A_n) \cup \{ s \}$.

3. If $A$ is a **dilemma resolving argument** $A_1, \ldots, A_n \Rightarrow s_1 > s_2$ then:
   - $\text{Prem}(A) = \text{Prem}(A_1) \cup \ldots \cup \text{Prem}(A_n)$,
   - $\text{Conc}(A) = s_1 > s_2$,
   - $\text{Sub}(A) = \text{Sub}(A_1) \cup \ldots \cup \text{Sub}(A_n) \cup \{ A \}$,
   - $\text{Norms}(A) = \text{Norms}(A_1) \cup \ldots \text{Norms}(A_n) \cup \{ \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow s_1 > s_2 \}$,
   - $\text{TopNorm}(A) = \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow s_1 > s_2$.

Given a set of arguments, some of them might be in conflict. For instance, two obligation arguments may be in conflict if their conclusions are contradictory, meaning that not both of the obligations can be accepted when both arguments have the same priority. In terms of argumentation theory, we say that these two arguments
defeat each other. Meanwhile, when one argument defeat another argument, the latter can be defeated in turn by other arguments. So, in order to evaluate the status of arguments, one needs first to identify the defeat relation over the arguments.

In the setting of normative systems, there are four types of propositions: elements (called brute facts) of the context, institutional facts, obligations and permissions. As mentioned in Section 3, the notion of normative or moral dilemma is traditionally defined as an unresolved conflict between two incompatible obligations, e.g., $Op \land O\neg p$ in modal logic. In terms of formal argumentation in this paper, it is represented by the existence of multiple extensions. Syntactically, it means two arguments supporting incompatible obligations defeat each other, and no priority can be applied between them. Meanwhile, normative conflict is brought about by conflicting institutional facts, which may also result in multiple extensions. Normative conflicts may lead to moral dilemmas. For example, if one stakeholder believes that there has been blasphemy while another stakeholder does not (a normative conflict), the first stakeholder may deduce that we should alert the police while the second stakeholder does not (a moral dilemma). In addition, according to Pigozzi and van der Torre [2018], two permissive norms never conflict, and a permissive norm is not in conflict with a brute fact or an institutional fact. Based on these considerations, the notions of priority relation and defeat relation between arguments are defined as follows.

Concerning the priority over arguments, according to the normative theory introduced in Section 3, constitutive norms always override other kinds of norms (otherwise wishful thinking) and permissive norms override regulative norms (as

![Diagram of priority order over different types of norms and arguments.](image)

**Figure 3**: The priority order over different types of norms and arguments. An exiting arrow indicates a higher priority.
they encode exceptions). So, an institutional fact argument may defeat an obligation argument, a permission argument may defeat an obligation argument, a context argument may defeat an institutional fact argument or an obligation argument, but not vice versa. In addition, context arguments have the highest priority. This is illustrated in Figure 3 and specified in Definition 10.

**Definition 10** (Priority relation between arguments). Let $\mathcal{A}$ be $\arg(\mathcal{A}_\mathcal{T})$ or $\arg(\mathcal{A}_\mathcal{N})$. Let $\mathcal{A}^b, \mathcal{A}^c, \mathcal{A}^r, \mathcal{A}^p \subseteq \mathcal{A}$ be the sets of brute fact arguments, institutional fact arguments, obligation arguments and permission arguments, respectively. Given two arguments $A, B \in \mathcal{A}$, we use $A > B$ to denote that $A$ is preferred to $B$, and $A \not> B$ to denote that $A$ is not preferred to $B$. We have

- For $A \in \mathcal{A}^b$ and $B \in \mathcal{A}^c \cup \mathcal{A}^r$, $A > B$.
- For $A \in \mathcal{A}^c$ and $B \in \mathcal{A}^r$, $A > B$.
- For $A \in \mathcal{A}^p$ and $B \in \mathcal{A}^r$, $A > B$.

In addition to the priority relation between different types of arguments, there can be priority relation between the same types of arguments. For instance, one obligation argument may be preferred to another. Since the latter is context dependent, and considering this relation does not affect the main point of our approach for checking and resolving moral dilemma, we leave abstract the priority relation between the same types of arguments.

Next, we define what it means for arguments to attack and defeat each other.

**Definition 11** (Attacks and defeats). Let $\mathcal{A}$ be $\arg(\mathcal{A}_\mathcal{T})$ or $\arg(\mathcal{A}_\mathcal{N})$. For all $A, B \in \mathcal{A}$, $A$ attacks $B$, iff $\text{Conc}(A) \in \phi$ for some $B' \in \text{Sub}(B)$ and $\text{Conc}(B') = \phi$; $A$ defeats $B$ iff $A$ attacks $B$ and $B \not> A$.

The set of defeats over arguments in $\arg(\mathcal{A}_\mathcal{T})$ is denoted as $\text{Def}(\mathcal{A}_\mathcal{T})$. Then $AF(\mathcal{A}_\mathcal{T}) = (\arg(\mathcal{A}_\mathcal{T}), \text{Def}(\mathcal{A}_\mathcal{T}))$ is called an argumentation framework constructed from $\mathcal{A}_\mathcal{T}$ ([Dung, 1995]).

Let $\arg(\mathcal{A}_\mathcal{N}) = \bigcup_{s \in \mathcal{N}} \arg(\mathcal{A}_\mathcal{T}s)$ be the union of all arguments in the argumentation frameworks of a set of stakeholders $\mathcal{N}$. The set of defeats over arguments in $\arg(\mathcal{A}_\mathcal{N})$ is denoted as $\text{Def}(\mathcal{A}_\mathcal{N})$.

**Example 4.4.** Continue Example 4.3. Three individual argumentation frameworks of the stakeholders $L$, $H$ and $M$ are illustrated in Figure 4(a). We see that argument $A_2$ attacks argument $A_3$, because reporting the information collected from users is in conflict with protecting their privacy.
The capital letters $L$, $H$, $M$ and $\{H, M\}$ next to an argument refer to the stakeholder (or the set of stakeholders) who have contributed to the argument while $c$, $r$ and $p$ indicate an institutional argument, an obligation argument and a permission argument respectively.

In terms of the work of Dung [1995] in an argumentation framework, a set of collectively acceptable arguments is called an extension. A core notion supporting the definition of various extensions is admissible sets. Specifically, given an argumentation framework $AF = (Arg, Def)$, a set of arguments is admissible, if and only if it is conflict-free and it can defend each argument within the set. A set $E \subseteq Arg$ is conflict-free if and only if there exist no arguments $A$ and $B$ in $E$ such that $(A, B) \in Def$. Argument $A \in Arg$ is defended by a set $E \subseteq Arg$ (also called $A$ is acceptable with respect to $E$) if and only if for all $B \in Arg$, if $(B, A) \in Def$, then there exists $C \in E$ such that $(C, B) \in Def$. Based on the notion of admissible sets, some other extensions could be defined. Formally, we have the following definition from Dung [1995].

**Definition 12** (Argumentation semantics). Let $AF = (Arg, Def)$ be an argumentation framework, and $E \subseteq Arg$ a set of arguments.

- $E$ is admissible iff $E$ is conflict-free, and each argument in $E$ is defended by $E$.
- $E$ is a complete extension iff $E$ is admissible and each argument in $Arg$ that is defended by $E$ is in $E$.
- $E$ is a preferred extension iff $E$ is a maximal (w.r.t. set-inclusion) complete extension.
- $E$ is a grounded extension iff $E$ is the minimal (w.r.t. set-inclusion) complete extension.
E is a stable extension iff E is conflict-free, and for each argument A ∈ Arg \ E, there exists B ∈ E, such that (B, A) ∈ Def.

In this paper, we use σ ∈ {co, pr, gr, st} to indicate complete, preferred, grounded, and stable semantics.

Given AT_s = (NS_s, K) of each stakeholder s ∈ S, and a decision problem, we distinguish the following three ways to check whether there is a dilemma, and use the dilemma resolving norms to deal with the dilemma if needed:

1. First, we consider each normative system of each stakeholder independently. In this case, we construct argumentation frameworks for each object level stakeholder, to check whether there is a dilemma.\(^3\)

2. Second, we consider all arguments of the stakeholders together. In this case, we construct a single argumentation framework to check whether there is a dilemma,\(^4\) in which each argument consists of a set of norms from a single normative system.

3. Third, we put all normative systems together, and a unified argumentation theory to check whether there is a dilemma.

4. Fourth, we use the Jiminy to decide among the stakeholders the most competent for the dilemma.\(^5\)

Given an argumentation framework AF(AT) = (Arg(AT), Def(AT)) of an argumentation theory AT (either AT_s or AT_N) and an extension E ∈ σ(AF(AT)), let Obl(E) = {Conc(A) | A ∈ E \ Arg^{Obl}(AT)} be the set of obligations in the conclusions of E.

Definition 13 (Individual framework dilemma checking and resolving). Let DP = (css, DV) be a decision problem. Let S_0 be a set of stakeholders, and IF = {AF(AT_s) | s ∈ S_0} be a set of individual argumentation frameworks constructed from the argumentation theories of the set of stakeholders S_0, where AT_s = (NS_s, css). There is a dilemma at the first level with respect to DP and IF under an argumentation semantics σ, if there exist E ∈ σ(AF_s), E' ∈ σ(AF_{s'}) where AF_s, AF_s' ∈ IF for some s, s' ∈ S_0, such that (Obl(E_1) ∪ Obl(E_2)) \ DV is inconsistent with respect to the contrariness function ¬. Otherwise, for all s, s' ∈ S_0, for all E ∈ σ(AF_s), E' ∈ σ(AF_{s'}), if (Obl(E_1) ∪ Obl(E_2)) \ DV = ∅, then there is no dilemma at the first level. Otherwise, the dilemma is resolved at the first level.

\(^3\)Here we call all stakeholders except Jiminy “object level stakeholders”, meaning that Jiminy is located at a meta-level to provide priority relations between the stakeholders.

\(^4\)A stakeholder could be in conflict with respect to her own norms.

\(^5\)The source of the Jiminy priorities is domain specific. We assume that the set of norms of Jiminy is given.
Example 4.5. Continue Example 4.4. The argumentation frameworks of stakeholders L, H and M have extensions \( E_1 = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_1 \} \), \( E_2 = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_2 \} \) and \( E_3 = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_4 \} \) respectively.

Since \( K = css = \{ w_1, w_2, w_3, w_4 \} \), \( Obl(E_1) = \{ d_1 \} \), and \( Obl(E_2) = \{ d_2 \} \), there is a dilemma at the first level with respect to the DP described in Example 4.1 and the individual argumentation frameworks illustrated in Figure 4.4(a).

For the second level checking a dilemma, we first combine all argumentation frameworks of the set of stakeholders \( S_0 = \{ s_1, \ldots, s_n \} \) to a single argumentation framework \( AF_{S_0} \), called a combined argumentation framework.

**Definition 14** (Combined framework). Let \( IFF = \{ AF(A_Ts) \mid s \in S_0 \} \) be a set of argumentation frameworks constructed from the argumentation theories of the set of stakeholders \( S_0 \), where \( AF(A_Ts) = (Arg(A_Ts), Def(A_Ts)) \). An argumentation framework combined from all argumentation frameworks in \( IFF \) is defined as \( AF_{S_0} = (Arg_{S_0}, Def_{S_0}) \), where \( Arg_{S_0} = \bigcup_{s \in S_0} Arg(A_Ts) \), and \( Def_{S_0} = \{(A, B) \mid A \text{ defeats } B \text{ according to Definition 14}\} \).

**Proposition 4.6.** Given a set of individual argumentation frameworks \( IFF = \{ AF(A_Ts) \mid s \in S_0 \} \) at the first level, and a combined argumentation framework \( AF_{S_0} = (Arg_{S_0}, Def_{S_0}) \) at the second level, it holds that \( Def_{S_0} \supseteq \bigcup_{s \in S_0} Def(A_Ts) \).

**Proof.** If \( A \text{ defeats } B \text{ according to } (A, B) \in Def(A_Ts) \text{ for all } s \in S_0 \), \( A \) still defeats \( B \) when \( A \) and \( B \) are in \( Arg_{S_0} \) according to Definition 14. According to definition 14, \( (A, B) \in Def_{S_0} \). So, it holds that \( Def_{S_0} \supseteq \bigcup_{s \in S_0} Def(A_Ts) \). \( \square \)

**Definition 15** (Combined framework dilemma checking and resolving). Let \( DP = (css, DV) \) be a decision problem, and \( AF_{S_0} \) be the argumentation framework combined from all argumentation frameworks \( AF(A_Ts) = (Arg(A_Ts), Def(A_Ts)) \) of the set of stakeholders \( S_0 \), where \( AT_s = (NS_s, css) \), \( s \in S_0 \). There is a dilemma at the second level with respect to \( DP \) and \( AF_{S_0} \) under an argumentation semantics \( \sigma \), if there exist \( E_1, E_2 \in \sigma(AF_{S_0}) \), such that \( Obl(E_1) \cup Obl(E_2) \cap DV \) is inconsistent with respect to the contrariness function \( ^\parallel \). Otherwise, if for all \( E_1, E_2 \in \sigma(AF_{S_0}) \), \( Obl(E_1) \cup Obl(E_2) \cap DV = \emptyset \), then there is no dilemma at the second level. Otherwise, the dilemma is resolved at the second level.

**Example 4.7.** Continue Example 4.3. By combining three argumentation frameworks of stakeholders L, H and M in Figure 4.4(a), we get a combined argumentation framework illustrated in Figure 4.4(b), which has three complete extensions \( E_1 = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_1 \} \), \( E_2 = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_2 \} \) and \( E_3 = \{ A_{01}, A_{02}, A_{03}, A_{04}, A_4 \} \). Since \( Obl(E_1) = \{ d_1 \} \) and \( Obl(E_2) = \{ d_2 \} \), the dilemma still exists.
However, a moral dilemma can be resolved at the second level sometimes. Consider the following example by combining argumentation frameworks from three stakeholders.

**Example 4.8.** Continue Example [4.7] Assume that the stakeholder \( M \) in \( S \) is now replaced by another stakeholder \( Q \). Let \( R_Q = \{ w_4 \Rightarrow \epsilon_Q, d_3 \} \), namely if the manufacturer is doing business in Norway, then they should report (when encountered collecting information) the information concerning a potential critical danger to society. Let \( \text{Arg}^1_A = \{ A_{01}, A_{02}, A_{03}, A_{04}, B \} \), where \( B = A_{04} \Rightarrow \epsilon_Q, d_3 \). It holds that \( B \) defeats \( A_1 \). By combining the three argumentation frameworks from three stakeholders \( H, L, Q \), a combined argumentation framework is illustrated on the right side of Figure 5(a), which has one complete extension \( E_1 = \{ A_{01}, A_{02}, A_{03}, A_{04}, B, A_2 \} \). Since there is one complete extension containing an obligation argument, i.e., \( \text{Obl}(E_1) = \{ d_2 \} \), the dilemma is solved.

For the third level resolution of a moral dilemma, we combine all normative systems from a set of stakeholders and construct an integrated argumentation framework.

**Definition 16** (Integrated framework dilemma checking and resolving). Let \( DP = (css, DV) \) be a decision problem, and \( AT_{S_0} = (NS_{S_0}, K) \) a combined argumentation theory of \( S_0 \), where \( NS_{S_0} = (L, \{ R_s \}, \bar{\cdot}) \), in which \( R_{S_0} = \bigcup_{s \in S_0} R_s \). There is a dilemma at the third level with respect to \( DP \) and \( AF(AT_{S_0}) \) under an argumentation semantics \( \sigma \), if and only there exist \( E_1, E_2 \in \sigma(AF(AT_{S_0})) \), such that \( (\text{Obl}(E_1) \cup \text{Obl}(E_2)) \cap DV \) is inconsistent with respect to the contrariness function \( \bar{\cdot} \). Otherwise,
if for all \( E_1, E_2 \in \sigma(AF(AT_{S_0})) \), \((\text{Obl}(E_1) \cup \text{Obl}(E_2)) \cap DV = \emptyset\), then there is no dilemma at the third level. Otherwise, the dilemma is resolved at the third level.

**Proposition 4.9.** Given a combined argumentation framework \( AF_{S_0} = (\text{Arg}_{S_0}, \text{Def}_{S_0}) \) at the second level and an integrated argumentation framework \( AF(AT_{S_0}) = (\text{Arg}(AT_{S_0}), \text{Def}(AT_{S_0})) \) at the third level, it holds that \( \text{Arg}_{S_0} \subseteq \text{Arg}(AT_{S_0}) \) and \( \text{Def}_{S_0} \subseteq \text{Def}(AT_{S_0}) \).

**Proof.** According to Definition 9, for all \( A \in \text{Arg}_{S_0} \), \( A \) can also be constructed from the combined argumentation theory \( AT_{S_0} \), and therefore \( A \in \text{Arg}(AT_{S_0}) \). So, \( \text{Arg}_{S_0} \subseteq \text{Arg}(AT_{S_0}) \). On the other hand, for all \( (A, B) \in \text{Def}_{S_0} \), \( A \) and \( B \) are in \( \text{Arg}_{S_0} \) and therefore in \( \text{Arg}(AT_{S_0}) \). According to definition of the defeat relation between \( A \) and \( B \) in \( \text{Arg}_{S_0} \) does not change when \( A \) and \( B \) are considered in \( \text{Arg}(AT_{S_0}) \), and therefore \( (A, B) \in \text{Def}(AT_{S_0}) \). So, \( \text{Def}(AT_{S_0}) \subseteq \text{Def}(AT_{S_0}) \). \( \square \)

**Example 4.10.** Continue Example 4.4. By combining three normative systems \( L, H \) and \( M \), we get a combined argumentation theory. The corresponding argumentation framework of this argumentation theory is illustrated in Figure 4.4(c), which has three complete extensions that are the same as the ones in Example 4.7. The dilemma still exists.

Similar to the case of dilemma resolution at the second level, in some situations a dilemma at the second level can be resolved at the third level.

**Example 4.11.** Let us consider another example based on a language \( L' \) that extends \( L \) described in Example 4.2, by adding new elements \( e_1 \) and \( e_2 \) such that \( \tilde{e}_1 = \{a_1, e_2\} \) and \( \tilde{e}_2 = \{a_2, e_1\} \). Consider two argumentation theories from stakeholders \( X \) and \( Y \), respectively: \( R_X = \{w_1 \Rightarrow_X a_1, a_1 \Rightarrow_X e_1\} \), \( R_Y = \{w_2 \Rightarrow_Y a_2, a_2 \Rightarrow_Y e_2, a_1 \Rightarrow_Y e_1\} \), \( R_I = \{w_3 \Rightarrow L > H\} \); \( K_1 = \{w_1, w_2, w_3, w_4\} \). Let \( B_1 = (A_{01} \Rightarrow_X a_1) \), \( B_2 = (B_1 \Rightarrow_Y e_1) \), \( B_3 = (A_{02} \Rightarrow_Y a_2) \), \( B_4 = (B_3 \Rightarrow_Y e_2) \), \( B_5 = (B_1 \Rightarrow_Y e_1) \). There exists a dilemma both at the first level and the second level. For the former, the two argumentation frameworks illustrated on the left upper side of Figure 5(b) have different complete extensions that contain arguments with contradictory obligations. For the latter, the argumentation framework illustrated on the left lower side of Figure 5(b) has two extensions. Two of them contain contradictory obligations. However, at the third level, the dilemma does not exist. This is because the integrated argumentation framework \( AF(AT_{[X,Y]}) \) has three extensions: \( \mathcal{E}_1 = \{A_{01}, A_{02}, A_{03}, A_{04}, B_1, B_5\} \), \( \mathcal{E}_2 = \{A_{01}, A_{02}, A_{03}, A_{04}, B_2, B_3, B_5\} \) and \( \mathcal{E}_3 = \{A_{01}, A_{02}, A_{03}, A_{04}, B_5\} \), and only \( \mathcal{E}_2 \) contains an obligation argument.

After introducing dilemma resolving at levels 2 and 3, one may notice that a dilemma may be resolved by adding defeats at the second level, or by adding
arguments and defeats at the third level. If a dilemma cannot be resolved at the third level, we may consider to remove defeats from the argumentation frameworks at the second and the third levels by applying the meta norms \( R_J \) from Jiminy.

After combining object level norms from the stakeholders and the dilemma resolving norms from the Jiminy, we get a combined argumentation theory containing dilemma resolving norms.

**Definition 17** (Argumentation framework with dilemma resolving arguments). Let \( S = \{s_1, \ldots, s_n, J\} \) be a set of stakeholders, containing the Jiminy \( J \). Given \( AT_s = (NS_s, \mathcal{K}) \) for all \( s \in S \), a combined argumentation theory is \( AT_S = (NS_S, \mathcal{K}) \), where \( NS_S = (L, \neg, R_S) \), in which \( R_S = \bigcup_{s \in S} R_s \). An argumentation framework constructed from \( AT_S \) is denoted as \( AF(AT_S) = (\text{Arg}(AT_S), \text{Def}(AT_S)) \), which contain dilemma resolving arguments.

Given an argumentation framework with dilemma resolving arguments, in some of its extensions, there might be some arguments that are dilemma resolving arguments. When the priority relation contained in the dilemma resolving arguments of an extension is not compatible with the extension, it should be removed.

**Example 4.12.** Continue Example 4.10. After adding norms from the Jiminy to the combined argumentation theory of all stakeholders, we may construct an argumentation framework illustrated in Figure 6, where \( A_6 = A_4 \Rightarrow L > H \) and \( A_7 = A_5 \Rightarrow H > L \). Under complete extensions, it has three extensions: \( E_1 = \{A_01, A_02, A_03, A_04, A_1, A_3, A_4, A_6\} \), \( E_2 = \{A_01, A_02, A_03, A_04, A_2, A_4, A_6\} \) and \( E_3 = \{A_01, A_02, A_03, A_04, A_4, A_6\} \). Note that the extension \( E_2 \) is not compatible with the priority relation \( L > H \) contained in the argument \( A_6 \), in the sense that the defeat from argument \( A_2 \) to \( A_1 \) no longer holds, and therefore \( A_2 \) can not be accepted, which contradicts \( A_2 \) being accepted with respect to \( E_2 \).

To address the problem illustrated in Example 4.12, we use a two-stage approach to obtain the extensions of \( AF(AT_S) \), based on the approach introduced by Brewka 1994. First, we compute the set of extensions of the argumentation framework with dilemma resolving arguments without considering the priority information contained in dilemma resolving arguments. Then, get a set of reduced argumentation frameworks by considering this priority information and check the compatibility of each extension. Formally, we say that an extension is compatible with respect to the priority relation of the dilemma resolving arguments in the extension if and only if it is contained in one of the extension of the reduced argumentation framework. In the terms of Brewka 1994, an extension will survive if it can be reconstructed after the priority information from the dilemma resolving arguments is considered.
Definition 18 (Reduced argumentation framework). Let $AF(AT_S)$ be combined argumentation framework of all stakeholders $S = \{s_1, \ldots, s_n, J\}$. Under a semantics $\sigma$, for all $E \in \sigma(AF(AT_S))$, let $E^J \subseteq E$ be the set of all dilemma resolving arguments, and $\text{Conc}(E^J) = \{\text{Conc}(A) \mid A \in E^J\}$. For all $A, B \in \text{Arg}(AT_S)$, we say that $A$ is preferred to $B$ with respect to the priority information contained in $E$, denoted $A \succ^E B$ if and only if $\forall s \in \text{Stakeholder}(A)$, $\forall s' \in \text{Stakeholder}(B)$, $s > s' \in \text{Conc}(E^J)$. A reduced argumentation framework with respect to $E$ is defined as $AF^E = (\text{Arg}, \text{Def}^E)$, where

$$\text{Def}^E = \{(A, B) \in \text{Def} \mid (B, A) \in \text{Def} \land (A \succ E B)\}$$

(1)

Definition 19 (Compatibility). We say that the priority contained in $E^J$ is compatible with $E$ if and only if there exists $E' \in \sigma(AF^E)$ such that $E \subseteq E'$.

Definition 20 (Priority extension). Let $AF(AT_S)$ be combined argumentation framework of all stakeholders $S = \{s_1, \ldots, s_n, J\}$. We say that $E$ is a priority extension of $AF(AT_S)$ if and only if it is an extension of $AF(AT_S)$ and is compatible with respect to the priority information contained in the dilemma resolving arguments of $E$.

Example 4.13. Continue Example 4.12. It holds that the argumentation framework in Figure 6 has two priority extensions $E_1$ and $E_3$. The reduced argumentation framework that has extension $E_1$ is illustrated in Figure 4(d).
Definition 21 (Integrated framework + Jiminy dilemma checking and resolving). Let \( DP = (\text{css}, DV) \) be a decision problem, and \( AT_S = (NS_S, \mathcal{K}) \) a combined argumentation theory of \( S \), where \( NS_S = (\mathcal{L}, \neg, R_S) \), in which \( R_S = \bigcup_{s \in S} R_s \). There is a dilemma at the fourth level with respect to \( DP \) and \( AF(\text{AT}_S) \) under an argumentation semantics \( \sigma \), if and only if there exists two priority extensions \( E_1, E_2 \in \sigma(AF(\text{AT}_S)) \), such that \((\text{Obl}(E_1) \cup \text{Obl}(E_2)) \cap DV \) is inconsistent with respect to the contrariness function \( \neg \). Otherwise, if for all priority extensions \( E_1, E_2 \in \sigma(AF(\text{AT}_S)) \), \((\text{Obl}(E_1) \cup \text{Obl}(E_2)) \cap DV = \emptyset \), then there is no dilemma at the fourth level. Otherwise, the dilemma is resolved at the fourth level.

Besides using norms from the Jiminy for the integrated framework, it is also feasible to combine them with the individual frameworks and the combined framework. The details of this combinations are omitted. Finally, we end this section with the following proposition.

Proposition 4.14. Given a set of normative systems \( NS_s = (\mathcal{L}, \neg, R_s) \) where \( s \in S = \{s_1, \ldots, s_n, J\} \), and a decision problem \( DP = (\text{css}, DV) \), the Jiminy will have one of the following three possible answers: there is a dilemma at level \( i \), there is no dilemma at level \( i \), and the dilemma is resolved at level \( i \), where \( i = 1, 2, 3, 4 \).

Proof. According to Definitions 13, 15, 16 and 21 this proposition directly holds. \( \square \)

5 Explaining Jiminy choices

Explainability is the problem of how a human can understand the decisions made by someone else in a given context. Recently, methodologies, properties and approaches to explanations in artificial intelligence have been widely studied (Biran and Cotton, 2017). The ethical decisions or recommendations that Jiminy makes are explainable. Generating explanations for Jiminy’s choices is a feature of the argumentation approach we take to reach agreements among the stakeholders, since argumentation has “a unique advantage in transparently explaining the procedure and the results of reasoning” (Fan and Toni, 2015, p. 1).

What is an explanation? Miller 2019 discusses the desirable features of an explanation from a social science point of view. He states that explanations are contrastive, in the sense that people expect an explanation not only about why one event happened, but (also) about why another event did not happen instead. Explanations are selected in the sense that all the causes of an event are not expected to be offered, rather a selection of one or two causes are selected for inclusion in the explanation. Truth and likelihood matter for an explanation, but a full probabilistic analysis of the event is not expected. Lastly, explanations are social in the sense that
they are presented with regard to the informational state of the person expecting an explanation.

All of the desirable aspects of explanations can be implemented in Jiminy. Contrastive explanations can be attained by considering all the available options that have been passed on to Jiminy and comparing this set with the option Jiminy ends up recommending. If there is a dilemma at any level, the recommendations from each of the extensions in the dilemma can be offered as possibilities, with an explanation as to why a particular extension survived resolution. Social explanations can be attained by using argument-based dialogues to formalize the process of explanations [Walton, 2011; Čyras et al., 2016; Cocarascu et al., 2018].

To explain why and how a decision is made by Jiminy, we first need to identify an argument in the extension whose conclusion is the decision. Meanwhile, to explain why another decision was not taken, we need to identify an argument in an argumentation framework whose conclusion is that other decision, and use the defeat relation among arguments to explain why an argument supporting that other decision is rejected.

More specifically, regarding the decision that was made, when the argument supporting that decision is located by referring to the argumentation framework, one may explain that the argument can be accepted because all of its attackers were rejected, which was in turn because at least one attacker of each of its attackers was accepted, and so on. In the context of this paper, whether a decision is made depends not only on the interaction between arguments in a single argumentation framework, but also on the assessed level of the decision, and on whether Jiminy plays the role of ranking the stakeholders.

Consider Figures 4(c) and (d) again. In the argumentation framework $\mathcal{AF}(\mathcal{AT}_{L,H,M})$, the options “comply with the law” ($d_1$) and “report information that grossly endangers society” ($d_3$) are justified, while the option “protect the privacy of users” is rejected. The explanations are as follows.

**Explaining derivability in arguments.** “Comply with the law” ($d_1$) is the conclusion of argument $A_1$, which can be derived from the context “the manufacturer makes the smart speaker” ($w_1$) and one norm stating “If you have manufactured a device, the behavior of that device should comply with the law” ($w_1 \Rightarrow_L d_1$). “Report information that grossly endangers society” ($d_3$) is the conclusion of $A_3$, which can be derived from the context “the information collected grossly endangers society” ($w_3$) and one norm stating “Devices that contain information about a future event that grossly endangers society should report that information to the authorities” ($w_3 \Rightarrow_H d_3$).

**Explaining justification and rejection as a dialogue by referring to an argumentation graph.** Argument $A_1$ is accepted because it has no attacker since
the defeat from \( A_2 \) to \( A_1 \) is removed by applying the priority relation encoded by norm \( w_3 \Rightarrow L > H \) from the normative system of Jiminy comparing Figures 4(c) and 4(d). Argument \( A_3 \) is accepted because its only attacker \( A_2 \) is rejected, and this is because \( A_1 \) is accepted.

The interaction described above can be represented as a dialogue game or a discussion game. Readers may refer to Vreeswijk and Prakken (Vreeswijk and Prakken, 2000) and Booth et al. (Booth et al., 2018) for details.

There is some related work on argumentation frameworks and generating explanations in them. Fan and Toni (2015) argue that argumentation semantics are built to answer the question of which subsets of arguments are good rather than why a particular argument is good. They propose a semantics that specifically for generating relevant explanations. In an argumentation graph, several arguments can fully justify the inclusion of an argument \( A \) in the extension. However, sometimes just a subsection of these arguments, a so-called related extension, is enough to justify the inclusion of \( A \) in the extension. This semantics identifies different types of explanations, all defined in terms of the related admissibility of arguments. Fan and Toni (2015) also offer a comprehensive overview of work in argumentation concerned with the problem of building explanations. Sileno et al. (2014) consider an answer set implementation of generating explanations from arguments that also integrates probabilistic reasoning.

6 The interface between Jiminy and the autonomous system

How we integrate Jiminy with the agent depends on what type of moral agent we need to construct, or rather whether the agent itself has any moral reasoning capabilities apart from Jiminy. Following the work of Moor (2006), an artificial agent can be one of four different types of morally sensitive agent: ethical-impact agent, implicit ethical agent, explicit ethical agent and full ethical agent.

A full ethical agent is one that is able to reason ethically at a human level. Clearly, no such artificial agents exist at the moment, and it is uncertain whether they can exist (Etzioni and Etzioni, 2017).

An ethical-impact agent does not make any ethically sensitive decisions itself and does not necessarily operate in ethically sensitive situations. However, by virtue of replacing some human activities with the artificial agent, we change the “moral environment” in which the agent operates. For example, a decision aid system that assesses risks and recommends insurance policies would not itself be making ethical decisions. However, if the data that the system uses is biased in some way,
the system can propagate and even enhance this bias, thus making the world a less ethical place.

An implicit ethical agent does make ethically sensitive decisions or operates in an ethically sensitive context. However, the agent’s actions are constrained so that unethical outcomes are avoided. One example of this approach is Arkin’s ethical governor (Arkin et al., 2009), but there is also the work of Dennis et al. 2016. Dyrkolbotn et al. 2018 further refined the definition of implicit ethical agent to specify agents who make ethically sensitive decisions without using their autonomy, regardless of the level of autonomy they have. This means that the agent does not reason about what is right or wrong, but has its options externally labeled as right or wrong and can only choose from the second set.

An explicit ethical agent also makes ethically sensitive decisions or operates in an ethically sensitive context. Unlike the implicit ethical agent, the explicit ethical agent is able to use its own autonomy and reasoning abilities to distinguish ethical from unethical outcomes and actions. An example of such a system is the General Dilemma Analyzer of Anderson and Leigh Anderson 2014.

By coupling a Jiminy component with an agent that has no ethical reasoning abilities, we can create an implicit ethical agent. In such an integration, Jiminy serves as an “external labeler” of actions for the purpose of avoiding unethical outcomes. Effectively, Jiminy acts as an ethical governor, constraining actions not recommended by the argumentation reasoning engine based on the normative systems representing the stakeholder. Rather than having one stakeholder assess the actions of the agent, as is the case with Arkin’s ethical governor, the system automatically reaches agreement among all identified stakeholders for this purpose. Figure 7a illustrates such an implicit ethical agent created by assigning a Jiminy component to the role of an ethical governor.

![Image](image.png)

Figure 7: Integrating Jiminy in an agent

We assume that the agent has a knowledge base and sensors to reason about its
environment, as well as a planner to identify possible actions. Each set of possible actions are communicated to Jiminy, whose reasoning cycle is triggered only when Jiminy identifies actions or situations involving the agent as being morally sensitive.

Explicit ethical agents are able to engage in ethical reasoning, and possibly also develop their own moral theories. By virtue of design, particularly if the agent is learning its moral theory, the stakeholders cannot be certain what the agent ends up treating as moral behavior. However, for some agents, it would be important to make sure that certain ethically sensitive situations are not left entirely to the autonomous decision making of the agent. This is where Jiminy in the role of ethical advisor can be used, interfacing not directly with the agent’s planner, knowledge base and possibly sensors, but with the agent’s ethical reasoning engine (see Figure 2b). Having Jiminy as an advisor does not change the resulting behavior of the agent, in the sense that the agent remains an explicit ethical agent.

There are (at least) two roles that Jiminy can play as a moral advisor. The ethical reasoning engine of the agent can simply delegate certain moral decisions to Jiminy. This means that there are specified ethically sensitive situations in which the ethical reasoner alone makes ethical choices, and then there are other specified situations in which Jiminy acts as governor and constrains some of the agent’s decisions while the ethical reasoner is not engaged. By playing this advisory role, the agent behaves as an explicit ethical agent in some contexts, and as an implicit ethical agent in others.

Alternatively, the agent’s ethical reasoner, in specified situations, becomes an additional stakeholder in Jiminy, and Jiminy constrains the actions of the agent. Now, the resulting agent remains an explicit ethical agent because it is the agent’s own ethical reasoner that is always involved in the agent’s ethical decision making. The problem of how to interface the agent and Jiminy so as to have the ethical reasoner provide its own normative system depends heavily on the specific abilities of the agent, and is outside the scope of this work at present.

It should be mentioned that for both advisory and governor integrations, Jiminy never interacts directly with the environment (or users of the agent), only with the other agent components. For reasons already heavily discussed in the literature, we can consider the possibility of providing users with a Jiminy off switch that simply disengages Jiminy (Hadfield-Menell et al., 2017), with the result that none of the actions the agent passes on to Jiminy will be constrained.

Regardless of whether Jiminy is used as an advisor or as a governor, its reasoning cycle (illustrated in Figure 2) remains the same, and we focus on specifying the subcomponents of its normative system, its argumentation reasoning engine, and its explanation generation engine. In the rest of the paper, we will work with one running example to illustrate different aspects of these subcomponents.

Consider a smart speaker like Google Home, Amazon Alexa or Apple Siri. In order to be able to react to a user’s request, the assistant has to passively “listen
in” and temporarily store the audio information from its environment. The smart speaker may use past voice recordings to improve future performance. In these respects, the smart speaker acts like a surveillance device. Moreover, during the course of a criminal investigation, the smart speaker may have data available that could help clarify whether a crime took place. It is questionable at present whether such data can be used in this way. It is less clear whether a smart speaker should be used as a crime prevention device, i.e., when the speaker detects that a crime is being planned, should the speaker alert someone about it?

If a crime is committed against users of the device, those users may want the device to alert the police or store relevant data to help resolve the crime. However, if the users themselves are planning a crime, they would not want the device to “snitch” on them. “Snitching” can be considered as an immoral outcome for the owner. As desirable as it might be for society or even a manufacturer to use the full extent of a device’s ability to solve and prevent crimes, no one would buy a device that severely limits their own freedom in their own home. Furthermore, whenever data recorded on the device is transferred to a third party, that in itself can be considered as an ethically sensitive context.

Assume that the smart speaker registers that a crime is being planned at home by some members of the household. The crime might be murder, or an act of mass terrorism. What is the appropriate boundary between privacy and security? Should the smart speaker alert the household, the authorities, or the manufacturer that such an event has taken place? The three choices constitute an example of a moral dilemma. Each choice could be considered moral. However, they cannot all be executed at the same time.

Legally, being aware that a crime is being planned and not reporting it can constitute conspiracy, which is an illegal act. However, the smart speaker itself cannot be “aware”, and as a non-person, is not criminally liable. Informing all members of the household of the criminal plans can cause those members to become liable to conspiracy charges. In the absence of such an alert, these charges could still be made since the household members have access to the audio files of the smart speaker. If the smart speaker brings the criminal activities to the attention of the entire household, it can help them choose whether to report the crime and thus isolate themselves from charges.

The manufacturer may choose to put the privacy of its customers above potential crime prevention. Not choosing to do so may mean losing customers. However, the manufacturer may not produce a device that breaks the law. But if the manufactured device reports on people continuously, such a device would not be sold. In some professions that are subject to client confidentiality, confidentiality can be broken

[https://www.theinformation.com/articles/amazon-echo-and-the-hot-tub-murder?](https://www.theinformation.com/articles/amazon-echo-and-the-hot-tub-murder?)
when clients are a threat to themselves or others.

Clearly, some crimes are more grievous than others. In a case where mass terrorism is planned, national security may be considered more important than privacy, and the norms coming from a stakeholder representing social safety on how to handle the situation should take precedence. Also, if this is the type of crime planned, household members would like to know. But what counts as mass terrorism and who decides this? Furthermore, assume that the smart speaker overheard an event such as blasphemy or domestic violence. Some countries count blasphemy as violence, others do not. Domestic violence is criminal in some countries and not in others. Both blasphemy and domestic violence are more “private” than murder and terrorism. In such cases, should the choice of the manufacturer or household be given precedence over the law? This is not a question that we will answer by the end of the paper, rather we build a system in which all options can be specified.

7 Related work

We distinguish related research in formal argumentation about normative systems from research in machine ethics and explainable AI. Concerning the former, in this paper we use only relatively abstract theories, because we believe that it is precisely this generality that makes the combination of normative systems and formal argumentation suitable for the Jiminy advisor. For a general background on these formal theories, see: the Handbook of Deontic Logic and Normative Systems (Gabbay et al., 2013), in particular the chapter on moral dilemmas by Lou Goble; the Handbook of Normative Multiagent Systems (Chopra et al., 2018); the Handbook of Formal Argumentation (Baroni et al., 2018a); and the formal argumentation manifesto (Gabbay et al., 2018). For an overview of the application of formal argumentation to normative systems, see the work of da Costa et al. (2018). The work of Arisaka et al. (2017) studies multi-agent argumentation at the abstract level, and the work of Pigozzi and van der Torre (2018) introduce a structured argumentation theory with constitutive and regulative norms. As far as we know, this paper is the first in the area of structured argumentation that considers moral dilemmas emanating from multiple normative systems representing several stakeholders.

To position the theory of normative systems and formal argumentation in the general area of knowledge representation and reasoning, it may be observed that both theories have been built on the Tarskian theory of deductive systems, i.e., mathematical proof theories in deductive logic, but they have also been built as criticisms of that theory. The main criticism of classical logic is the monotonicity property, and these two theories can be rephrased in the framework of nonmonotonic logic. They are typically concerned with both theoretical reasoning and practical
reasoning. There are many distinct versions of theories of normative systems as well as many distinct theories of formal argumentation. These knowledge representation and reasoning formalisms have been used in many disciplines. Consequently, there are relatively abstract theories that can be used across disciplines, and more detailed theories developed to be used in specific disciplines because they have been adapted to the specific concerns of those disciplines.

Our definition of argumentation theory conforms to the abstract language used in ASPIC+ (Modgil and Prakken 2013) and some other work that extend ASPIC+, particularly by Baroni et al. 2015; 2018b, where the contrariness function is used. However, compared to ASPIC+, the definition of argumentation theory in this paper is somewhat simpler: since we assume that all norms are defeasible, we use only defeasible rules. Meanwhile, we do not deal with domain dependent priorities over rules. However, in order to adapt to the different types of norms, we use three kinds of rules to represent institutional norms, regulative norms and permissive norms respectively. In addition, since we only use defeasible rules, the problem of the contrariness function mentioned by Baroni et al. 2015; 2018b does not exist.

Secondly, concerning the definition of a defeat relation, we only use rebut, and it is sufficient to model the conflicting relation between norms. For the priority relation over arguments, the conflicts between different types of norms have different properties, i.e., two permissive norms are never in conflict, institutional norms are preferred to permissive norms, and permissive norms are preferred to regulative norms, and so we provide a domain independent definition of priority over different kinds of arguments. This differs from some other work involving prioritized argumentation. For instance, Young et al. 2016 and Liao et al. 2016 use prioritized argumentation to represent different kinds of prioritized nonmonotonic formalisms like Reiter’s default logic (Reiter, 1980) and Brewka and Eiter’s Preferred Answer Sets (Brewka and Eiter, 1999), but they do not focus on how to represent the normative reasoning in terms of different types of norms.

Thirdly, with regards to reasoning about preferences in argumentation frameworks, Modgil 2009 proposes an approach that extends Dung’s theory to accommodate arguments that claim preferences among other arguments. Our work on accommodating the dilemma of resolving norms to the argumentation is in line with this work. In this paper, for simplicity, we did not apply the semantics of Modgil’s extended argumentation framework (Modgil 2009). Instead, we used a two-stage approach to obtain the extensions of an integrated argumentation framework, based on the approach introduced by Brewka 1994.

Fourthly, there is also interesting work about exploiting argumentation to model moral reasoning. For instance, Bench-Capon and Modgil 2017 propose an approach using an argumentation scheme based on values and designed for practical reasoning, and they show how this reasoning can be used to think about situations when norms
should be violated. Atkinson et al. [2018] continue this line of work, and present an approach to taking the actions of others into account based on argumentation schemes and value-based reasoning. We did not use argumentation schemes and value-based reasoning in our work. Instead, ASPIC+ style formal argumentation is used to model the dilemma checking and to resolve cases where a set of stakeholders have different opinions represented by a set of norms.

Concerning work on machine ethics, there is no consensus on whether an artificial agent can ever be a moral agent as categorically as people are (Moor, 2006; Etzioni and Etzioni, 2017). It is widely accepted that some level of moral behavior can be implemented in machines. Wallach and Allen [2008] distinguish between operational morality, functional morality, and full moral agency. Moor [2006] distinguishes between ethical-impact, explicit ethical, implicit ethical and full ethical agency; see also the work of Dyrkolbotn et al. [2018]. Some proposals and prototypes on how to implement moral agency are already being put forward, such as those of Anderson and Leigh Anderson [2014], Arkin et al. [2012], Bringsjord et al. [2008], Vanderelst and Winfield [2018], Dennis et al. [2016], and Lindner and Bentzen [2017].

It has been shown that people consider that the same ideas of morality do not apply to both people and machines (Malle et al., 2015). It is argued by Charisi et al. [2017] that the complex issue of where machine morality comes from should be considered from the aspect of all stakeholders—all the people who are in some way impacted by the behavior and decisions of an autonomous system. They distinguish government and societal regulatory bodies from manufacturers and designers and again from end users, customers and owners. Note that these broad categories of stakeholders can further be subdivided. For example, owners can be distinguished from “leases” of the autonomous system[7]. While it has been argued in the literature (Dignum [2017], Charisi et al. [2017]) that an autonomous system should be built to integrate moral, societal and legal values, to the best of our knowledge, no approach has been proposed on how to accomplish this. This paper is the first work that explicitly considers the problem of integrating the moral values of multiple stakeholders in an artificial moral agent.

The EU General Data Protection Regulation (GDPR), specifically Sections 13–15, gives users affected by automated decision making the right to obtain “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”. One way of obtaining this is by building systems capable of giving arguments to support the decisions they make. Our approach provides a way to do this.

[7]https://robohub.org/should-a-carebot-bring-an-alcoholic-a-drink-poll-says-it-depends-on-who-owns-the-robot/
Explainability has not been considered as a critical feature in logic-based systems—see for example the work of Dennis et al. [2016], Lindner and Bentzen [2017], and Bringsjord et al. [2008]. This is because one can use formal methods to prove what kind of behavior is possible for an autonomous systems in which contexts. We argue, however, that a formal proof, while “accessible” to a regulatory body, is not enough to constitute explainability for common people. The GenEth system (Anderson and Leigh Anderson [2014]) uses input from professional ethicists and machine learning to create a principle of ethical action preference. GenEth can “explain” its decisions with reference to how two options were compared and the ethical features of each option.

8 Summary

This paper proposes a Jiminy advisor for autonomous agents. Jiminy is a multiple-stakeholder ethical advisory component based on a theory of normative systems and formal argumentation. A knowledge engineer elicits the normative systems of the stakeholders, which may be viewed as tables. These are used to classify situations in terms of a set of ethically relevant features, and relate these features to normative decisions. The normative systems are represented efficiently as sets of constitutive and regulative norms, including permissive norms to represent exceptions. The argumentation system is a reasoning engine dedicated to finding moral agreements.

In the initial state, no consideration is given to interaction among the normative systems of the stakeholders. Each normative system is treated independently, and the advice of all the stakeholders are compared. Where there is disagreement about the deontic decision, for example when some of the stakeholders advise alerting the police while other stakeholders do not support this action, then we classify the situation as a moral dilemma. In such cases of moral dilemma, the argumentation engine proceeds in three steps.

First, the argumentation engine considers the combination of all the arguments of the stakeholders. At the abstract level, this means that attack relations among the arguments are taken into account. Instead of an argumentation framework for each stakeholder, now there is a large framework consisting of all the arguments of the stakeholders, together with the attack relations. If this leads to only one possible decision, then there is moral agreement and Jiminy returns that decision.

Second, where the dilemma is not resolved by combining the argumentation frameworks, then Jiminy will combine the three normative systems into a single normative system. As a consequence, there can be new arguments built from norms of distinct stakeholders, and the combined knowledge may be sufficient to reach moral agreement.
Third, and only where these two other methods have failed, Jiminy considers its stakeholder selection norms. These meta-norms are context dependent norms that select one stakeholder whose expertise is the most relevant. The effect of the stakeholder selection norms is to remove attacks on the arguments of the most relevant stakeholder originating from the arguments of other stakeholders.

It has often been observed that a major advantage of formal argumentation is that the reasoning process can be represented as a graph in which the nodes represent abstract arguments and the edges represent abstract relations between the arguments. The Jiminy architecture extends this approach to abstract analysis to resolving moral dilemmas among stakeholders. In the first step, attacks are added among the arguments of stakeholders; in the second step, arguments are added to the argumentation framework; in the third step, attack relations are removed from the framework.

This abstract representation of the resolution of moral dilemmas plays a central role in the explanation module of the Jiminy advisor. Besides the logical analysis of the derivability of an institutional fact or deontic conclusion within an argument, we can use techniques from abstract argumentation such as interactive dialogue procedures.

In future work, we can use formal methods from knowledge representation and reasoning or social choice to further study elements of specific normative systems, such as their conflict freeness or completeness, or specific argumentation parameters such as argumentation semantics. These formal analyses will make the workings of Jiminy even more transparent.

Moreover, our model of multi-stakeholder agreement can also be considered for other domains, such as the law. In international law, each country is assumed to be autonomous, and it is assumed that there is no ranking between countries. Nevertheless, sometimes incidents can concern various countries, particularly in inheritance or contracting matters. Thus, the relation between countries is analogous to the relation between the stakeholders in Jiminy. One difference between our ethical advisor and an international law advisor is that Jiminy has a single normative system for stakeholder selection whereas in international law, each national law contains a legal code to decide what is to be done in cross-border incidents. Another question is whether existing solutions in the law can also be used to further develop the ethical advisor introduced in this paper.

Acknowledgment A shorter version of this paper was presented at the In AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, January 27-28, 2019. That paper presents the general idea of the Jiminy advisor, but not the details of the formal argumentation mechanism. The authors are grateful to Dr. Louise
Dennis for providing some of the examples in this paper. Beishui Liao was partially supported by the Fundamental Research Funds for the Central Universities of China for the project Big Data, Reasoning and Decision Making, and the National Social Science Foundation of China (No.18ZDA290, No.17ZDA026). Leendert van der Torre acknowledges financial support from the Fonds National de la Recherche Luxembourg (INTER/Mobility/19/13995684/DLAI/van der Torre).

References

Alchourron, C. E., 1991. Conflicts of norms and revision of normative systems. Law and Philosophy 10, 413–425.

Alchourrón, C. E., Bulygin, E., 1981. The expressive conception of norms. In: Hilpinen, R. (Ed.), New studies in deontic logic. D. Reidel, Dordrecht, pp. 95–124.

Anderson, M., Leigh Anderson, S., 2014. GenEth: A general ethical dilemma analyzer. In: Proceedings of the 28th AAAI Conference on AI. pp. 253–261.
URL http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8308

Arisaka, R., Satoh, K., van der Torre, L. W. N., 2017. Anything you say may be used against you in a court of law - abstract agent argumentation (triple-a). In: AI Approaches to the Complexity of Legal Systems - AICOL International Workshops 2015-2017: AICOL-VI@JURIX 2015, AICOL-VII@EKAW 2016, AICOL-VIII@JURIX 2016, AICOL-IX@ICAIL 2017, and AICOL-X@JURIX 2017, Revised Selected Papers. pp. 427–442.
URL https://doi.org/10.1007/978-3-030-00178-0_29

Arkin, R., Ulam, P., Duncan, B., 2009. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Tech. Rep. GIT-GVU-09-02, Mobile Robot Laboratory, College of Computing, Georgia Tech.

Arkin, R., Ulam, P., Wagner, A. R., 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. Proc. of the IEEE 100 (3), 571–589.
URL http://dx.doi.org/10.1109/JPROC.2011.2173265

Atkinson, K., Bench-Capon, T. J. M., 2018. Taking account of the actions of others in value-based reasoning. Artif. Intell. 254, 1–20.
URL https://doi.org/10.1016/j.artint.2017.09.002

44
Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J. and Shariff, A., Bonnefon, J., Rahwan, I., 2018. The moral machine experiment. Nature 563 (7729), 59–64. URL https://doi.org/10.1038/s41586-018-0637-6

Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (Eds.), 2018a. Handbook of Formal Argumentation. College Publications.

Baroni, P., Giacomin, M., Liao, B., 2015. Dealing with generic contrariness in structured argumentation. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015. pp. 2727–2733. URL http://ijcai.org/Abstract/15/386

Baroni, P., Giacomin, M., Liao, B., 2018b. A general semi-structured formalism for computational argumentation: Definition, properties, and examples of application. Artif. Intell. 257, 158–207. URL https://doi.org/10.1016/j.artint.2018.01.003

Baum, S. D., 2020. Social choice ethics in artificial intelligence. AI & SOCIETY 35 (1), 165–176. URL https://doi.org/10.1007/s00146-017-0760-1

Bench-Capon, T. J. M., Modgil, S., 2017. Norms and value based reasoning: justifying compliance and violation. Artif. Intell. Law 25 (1), 29–64. URL https://doi.org/10.1007/s10506-017-9194-9

Bench-Capon, T. J. M., Prakken, H., Sartor, G., 2010. Argumentation in Legal Reasoning. Argumentation in Artificial Intelligence. Springer.

Biran, O., Cotton, C., 2017. Explanation and justification in machine learning: A survey. In: Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence (XAI 2017). pp. 8–13.

Bjørgen, E. P., Madsen, S., Bjørknes, T. S., Heimsæter, F. V., Hávik, R., Linderud, M., Longberg, P., Dennis, L. A., Slavkovik, M., 2018. Cake, death, and trolleys: Dilemmas as benchmarks of ethical decision-making. In: Furman, J., Marchant, G. E., Price, H., Rossi, F. (Eds.), Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018. ACM, pp. 23–29. URL https://doi.org/10.1145/3278721.3278767

Boella, G., van der Torre, L., 2006. Constitutive norms in the design of normative multiagent systems. In: Computational Logic in Multi-Agent Systems, 6th International Workshop, CLIMA VI. LNCS 3900. Springer. pp. 303–319.
Booth, R., Caminada, M., Marshall, B., 2018. DISCO: A web-based implementation of discussion games for grounded and preferred semantics. In: COMMA. Vol. 305 of Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 453–454.

Brandt, F., Conitzer, V., Endriss, U., Lang, L., Procaccia, A. D. (Eds.), 2016. Handbook of Computational Social Choice. Cambridge University Press. URL https://doi.org/10.1017/CBO9781107446984

Bremner, P., Dennis, L., Fisher, M., Winfiled, M., 2019. On proactive, transparent and verifiable ethical reasoning for robots.

Brewka, G., 1994. Reasoning about priorities in default logic. In: Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 2. pp. 940–945. URL http://www.aaai.org/Library/AAAI/1994/aaai94-144.php

Brewka, G., Eiter, T., 1999. Preferred answer sets for extended logic programs. Artif. Intell. 109 (1-2), 297–356. URL https://doi.org/10.1016/S0004-3702(99)00015-6

Bringsjord, S., Arkoudas, K., Bello, P., 2008. Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems 21 (4), 38–44.

Broersen, J. M., Dastani, M., Hulstijn, J., Huang, Z., van der Torre, L., W. N., 2001. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In: Agents. pp. 9–16. URL http://portal.acm.org/citation.cfm?id=375735.375766

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A., Yampolskiy, R., 2017. Towards moral autonomous systems. CoRR abs/1703.04741. URL http://arxiv.org/abs/1703.04741

Chopra, A., van der Torre, L., Verhagen, H., Villata, S., 2018. Handbook of normative multiagent systems. College Publications.

Cocarascu, O., Čyras, K., Toni, F., 2018. Explanatory predictions with artificial neural networks and argumentation. In: Proceedings of the IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI 2018). pp. 26–32.

da Costa Pereira, C., Liao, B., Malerba, A., Rotolo, A., Tettamnzi, A. G. B., van der Torre, L., Villata, S., 2018. Handling norms in multi-agent system by means of formal argumentation. In: Chopra, A., van der Torre, L., Verhagen, H., Villata,
S. (Eds.), Handbook of normative multiagent systems. College Publications, pp. 345–373.

Dennis, L. A., Fisher, M., Slavkovik, M., Webster, M. P., 2016. Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems 77, 1–14.
URL http://www.sciencedirect.com/science/article/pii/S0921889015003000

Dignum, V., 2017. Responsible autonomy. In: Proceedings of the 26th IIJCAI. pp. 4698–4704.
URL https://doi.org/10.24963/ijcai.2017/655

Dung, P. M., 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77 (2), 321–358.

Dyrkolbotn, S., Pedersen, T., Slavkovik, M., 2018. On the distinction between implicit and explicit ethical agency. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018. pp. 74–80.
URL https://doi.org/10.1145/3278721.3278769

Etzioni, A., Etzioni, O., 2017. Incorporating ethics into artificial intelligence. The Journal of Ethics 21 (4), 403–418.
URL https://doi.org/10.1007/s10892-017-9252-2

Fan, X., Toni, F., 2015. On computing explanations in argumentation. In: AAAI. AAAI Press, pp. 1496–1502.

Gabbay, D., Horyt, J., Parent, X., van der Meyden, R., van der Torre, L. (Eds.), 2013. Handbook of Deontic Logic and Normative Systems. College Publications, London.

Gabbay, D. M., Giacomin, M., Liao, B., van der Torre, L. W. N., 2018. Present and future of formal argumentation (dagstuhl perspectives workshop 15362). Dagstuhl Manifestos 7 (1), 69–95.
URL https://doi.org/10.4230/DagMan.7.1.69

Grossi, D., Jones, A., 2013. Constitutive norms and counts-as conditionals. In: Horyt, J., Gabbay, D., Parent, X., van der Meyden, R., van der Torre, L. (Eds.), Handbook of Deontic Logic and Normative Systems. College Publications, London, pp. 407–441.
Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S., 2017. The off-switch game. In: The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA. URL http://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15156

Horty, J. F., 1994. Moral dilemmas and nonmonotonic logic. J. Philos. Log. 23 (1), 35–65. URL https://doi.org/10.1007/BF01417957

Liao, B., Oren, N., van der Torre, L. W. N., Villata, S., 2016. Prioritized norms and defaults in formal argumentation. In: Roy, O., Tammininga, A. M., Willer, M. (Eds.), Deontic Logic and Normative Systems - 13th International Conference, DEON 2016, Bayreuth, Germany, July 18-21, 2018. College Publications, pp. 139–154.

Lindahl, L., Odelstad, J., 2013. TJS. a formal framework for normative systems with intermediaries. In: Horty, J., Gabbay, D., Parent, X., van der Meyden, R., van der Torre, L. (Eds.), Handbook of Deontic Logic and Normative Systems. College Publications.

Lindner, F., Bentzen, M., 2017. The hybrid ethical reasoning agent IMMANUEL. In: Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017. pp. 187–188. URL http://doi.acm.org/10.1145/3029798.3038404

Makinson, D., 1999. On a fundamental problem of deontic logic. In: McNamara, P., Prakken, H. (Eds.), Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science. IOS Press, pp. 29–54.

Makinson, D., van der Torre, L., 2000. Input-output logics. Journal of Philosophical Logic 29 (4), 383–408.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C., 2015. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. HRI '15. ACM, pp. 117–124. URL http://doi.acm.org/10.1145/2696454.2696458

Martino, A. A., Socci Natali, F. (Eds.), 1986. Permissive Norms and Normative Concepts. Amsterdam: North Holland.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38.
Modgil, S., 2009. Reasoning about preferences in argumentation frameworks. Artif. Intell. 173 (9-10), 901–934. URL https://doi.org/10.1016/j.artint.2009.02.001

Modgil, S., Prakken, H., 2013. A general account of argumentation with preferences. Artif. Intell. 195, 361–397. URL https://doi.org/10.1016/j.artint.2012.10.008

Moor, J. H., Jul. 2006. The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems 21 (4), 18–21.

Pigozzi, G., van der Torre, L., 2018. Arguing about constitutive and regulative norms. Journal of Applied Non-Classical Logics 28 (2-3), 189–217. URL https://doi.org/10.1080/11663081.2018.1487242

Reiter, R., 1980. A logic for default reasoning. Artif. Intell. 13 (1-2), 81–132. URL https://doi.org/10.1016/0004-3702(80)90014-4

Robinson, P., 2021. Moral disagreement and artificial intelligence. In: Fourcade, M., Kuipers, B., Lazar, S., Mulligan, D. K. (Eds.), AIES ’21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021. ACM, p. 209. URL https://doi.org/10.1145/3461702.3462534

Ross, A., 1941. Imperatives and logic. Theoria 7, 53–71, Reprinted in Philosophy of Science 11:30–46, 1944.

Ross, A., 1957. Tû-tû. Harvard Law Review 70, 812?825.

Searle, J., 1969. Speech Acts. An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

Shariff, A., Bonnefon, J., Rahwan, I., 2017. Psychological roadblocks to the adoption of self-driving vehicles. Nature Human Behaviour 1 (10), 694–696. URL https://doi.org/10.1038/s41562-017-0202-6

Sileno, G., Boer, A., van Engers, T., 2014. Implementing explanation-based argumentation using answer set programming. In: Eleventh International Workshop on Argumentation in Multi-Agent Systems (ARGMAS 2014): Paris, France, May 5, 2014: in conjunction with AAMAS 2014. http://hdl.handle.net/11245/1.430613, Cambridge, MA: Massachusetts Institute of Technology.

Stenius, E., 1963. Principles of a logic of normative systems. Acta Philosophica Fennica 16, 247–260.
Vanderelst, D., Winfield, A. F. T., 2018. An architecture for ethical robots inspired by the simulation theory of cognition. Cogn. Syst. Res. 48, 56–66. URL https://doi.org/10.1016/j.cogsys.2017.04.002

Čyras, K., Satoh, K., Toni, F., 2016. Explanation for case-based reasoning via abstract argumentation. In: Computational Models of Argument - Proceedings of COMMA. pp. 243–254. URL https://doi.org/10.3233/978-1-61499-686-6-243

Vreeswijk, G., Prakken, H., 2000. Credulous and sceptical argument games for preferred semantics. In: JELIA. Vol. 1919 of LNCS. Springer, pp. 239–253.

Wallich, W., Allen, C., 2008. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.

Walton, D., 2011. A dialogue system specification for explanation. Synthese 182 (3), 349–374.

Young, A. P., Modgil, S., Rodrigues, O., 2016. Prioritised default logic as rational argumentation. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016. pp. 626–634. URL http://dl.acm.org/citation.cfm?id=2937018