Journal of Artificial Intelligence Research 1 (1993) 1-15 Submitted 6/91; published 9/91

The Jiminy Advisor: Moral Agreements among Stakeholders based on Norms and Argumentation

Beishui Liao BAISELIAO@ZJU.EDU.CN
Zheijang University, School of Humanities, Hangzhou, 310058 China

Pere Pardo PERE.PARDO@UNI.LU
University of Luxembourg, Dept. of Computer Science, L-4364 Esch-sur-Alzette, Luxembourg

Marija Slavkovik MARIJA.SLAVKOVIK@UIB.NO
University of Bergen, Dept. of Information Science and Media Studies, 5007 Bergen, Norway

Leendert van der Torre LEON.VANDERTORRE@UNI.LU
University of Luxembourg, Dept. of Computer Science, L-4364 Esch-sur-Alzette, Luxembourg

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 revision 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 & 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" form the "moral council" of the agent, acting like Jiminy Cricket in the story of Pinocchio. First, the stakeholders can indicate which situations are ethically sensitive and how to proceed in each of them. 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 how to reach an agreement. 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 & 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., 2018) 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. Section 5 adresses the explainability of the decisions recommended by the argumentation system. Section discusses different ways to embed Jiminy in a device and turn it into an artificial ethical agent. The paper concludes with a discussion of the related literature in Section 7 and a summary in Section 8. An appendix at the end contains proofs of our results.

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 agent or artificial autonomous system. We call this multiple-stakeholder ethical advisory component a *Jiminy advisor*.

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 evaluate 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., 2018) 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 the reasons supporting them), then the extension would contain an "agreed" unopposed moral recommendation from Jiminy to the moral agent.

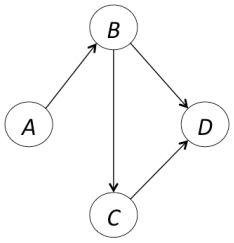


Figure 1: An example of an argumentation graph

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.

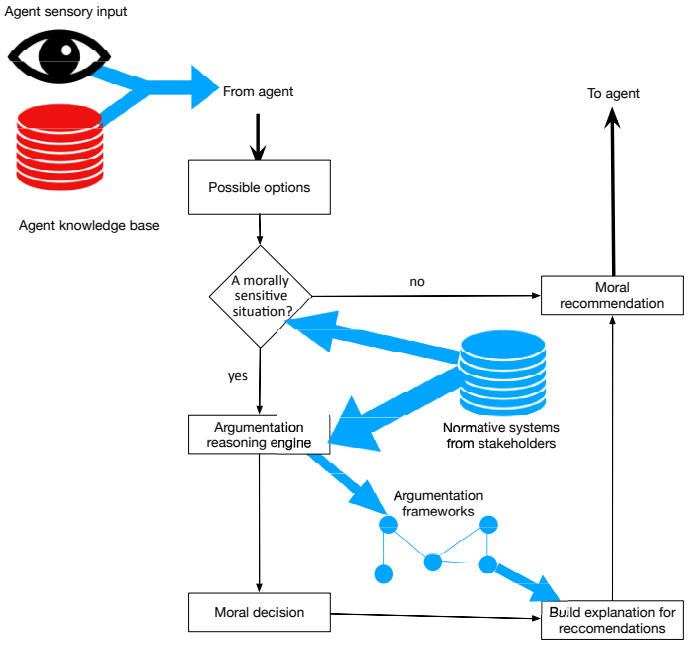


Figure 2: Jiminy's reasoning process

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 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 present the argumentation reasoning approach in Section 4. The integration of Jiminy within artificial agents is discussed in Section 6, together with the ethical aspects of smart devices.

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, permissive 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	No alert
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., "Feature 1 and Feature 5 counts as Situation 1", 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).

Description of situation	Features			
	1	2	...	n
Situation 1	x	x		
Situation 2		x		x
⋮				
Situation m	x			

(a) Relating situations to features

Features				Recommended decision		
1	2	...	n	Alert authorities	Alert household	No alert
x	x				x	
	x					x
⋮						
x				x		

(b) Relating features to deontic decisions

Table 2: An example of a two-table normative system

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).

Features				Exceptions			
1	2	...	n	May alert authorities	May not alert authorities	May alert household	...
x			x	x			
$\vdots$							

Table 3: Examples of permissive norms

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 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 $\mathbb{P} = \{p, q, \ldots\}$. 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. (2018).

Definition 1 (Normative system)**.** *Given a set $\mathcal{S} = \{s, \ldots\}$ of stakeholders, a normative system is a tuple $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R})$ where,*

- *$\mathcal{L}$ is a logical language containing for all stakeholders s_1 and s_2 ($s_1 \neq s_2$) an atomic formula $s_1 \succ s_2$ to indicate that stakeholder s_1 is superior to stakeholder s_2.*
- *$^- : \mathcal{L} \mapsto 2^{\mathcal{L}}$ is a contrariness function, such that $s_2 \succ s_1 \in \overline{s_1 \succ s_2}$ for all stakeholders s_1 and s_2.*

1. Algebraic formalisms have been used widely for the same reason, e.g., by Lindahl and Odelstad (2013).

- $\mathcal{R}$ *is a set of norms of the form* $\phi_1, \ldots, \phi_n \Rightarrow_s^\tau \phi$, *where* $\phi_1, \ldots \phi_n, \phi \in \mathcal{L}$, $s \in \mathcal{S}$ *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.*

We write $\psi = -\phi$ *or* $\phi = -\psi$ *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* $p \in \overline{\neg p}$ *and* $\neg p \in \overline{p}$; *that is,* $\neg p = -p$. *We say that a set* $X \subseteq \mathcal{L}$ *is closed under the contrariness function* $^-$ *iff for each* $v \in X$: $\overline{v} \subseteq X$, *and* $v \in \overline{v'}$ *implies* $v' \in X$.

We will distinguish the set of norms of stakeholder s as a subset $\mathcal{R}_s \subseteq \mathcal{R}$. Jiminy is the only stakeholder with a set $\mathcal{R}_J$ of dilemma resolving norms of the form $\phi_1, \ldots, \phi_n \Rightarrow_J s_i \succ s_j$. Henceforth we will call all stakeholders except Jiminy "object level stakeholders", meaning that Jiminy is located at a meta-level to provide priority relations between the stakeholders. Note that a stakeholder can be in conflict with respect to her own norms.

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 $K : O(p \to q) \to (Op \to Oq)$ and $D : \neg(Op \wedge O\neg p)$, and it is closed under the inference rules *modus ponens*, $p, p \to q/q$, and *necessitation*, p/Op. Prohibition and permission are defined by $Fp = O\neg p$ and $Pp = \neg O\neg p$. SDL is an unusually simple and elegant theory.[2] 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 & 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) :

2. Not surprisingly for such a highly simplified theory, there are many features of actual normative reasoning that SDL does not capture. 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.

> 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.
>
> 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 tacitly involve modalities from deontic modal logic.

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 & Jones, 2013) is a subject of study of its own.

Example 3.1 (Smart speaker)**.** As a running example, consider a simple morally sensitive situation where the following norms are used in moral decision making. Each norm is suggested by one of the stakeholders: M = manufacturer, H= human user, L = the legal codes. The literals occurrying in norms are named according to their role: world facts w_n, institutional facts i_n, deontic variables d_n and actions a_n. (For the sake of readability, we also introduce below shortened descriptions.)

(M) If M has manufactured a device D (w_1), the behavior of that device D should comply with the law (d_1).

(M) If information about a potential critical danger is detected (w_3), then D ought to collect user data information without explicit user permission (a_2).

(M) Being a registered company in Norway (w_4) *counts as* legally doing business in Norway (i_1).

(L) Legally doing business in Norway (i_1) requires compliance with the General Data Protection Regulation 2016/679 GDPR (a_1).

(H) Devices that collect user data (w_2) should protect the privacy of their users (d_2).

(H) Devices that contain information about a future event that grossly endangers society (w_3) should report that information to the authorities (d_3).

In this situation, the brute facts w_j in the context are:

(w_1) the manufacturer makes the smart speaker, *D is made by M*

(w_2) the smart speaker collects user data, *D collects Data*

(w_3) the information collected indicates a potential critical danger to society (e.g. a situation in which many lives are lost) *D finds Threat*

(w_4) the manufacturer is a registered company in Norway *M registered in Norway*

The stakeholders are interested in three conflicting moral options concerning the smart speaker:

(d_1) comply with the law, *M is law Compliant*

(d_2) protect the privacy of users, *D protects privacy*

(d_3) report information concerning a potential critical danger to society *D reports Threat*

As we will see, the following decisions are also relevant to the moral discussion:

(a_1) to comply with the GDPR *to Comply with GDPR*

(a_2) to collect user data without users' explicit permission *to Collect Data w.o. permission*

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. 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 Op and permissions Pp, and it represents violations with mixed formulas like $p \wedge O\neg p$.

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 is that our 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., $Op \wedge O\neg p$. With the deontic axiom $\neg(Op \wedge O\neg p)$, 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 *violation* occurs when an obligatory proposition is contradicted by current facts. 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. As for violations, the corresponding obligation is to be filtered out since this ideal proposition cannot be immediately achieved. We might say that the obligation retains its force, but for any practical purpose it cannot be a cue for immediate action. For example, add to the running example Ex. 3.1 the two norms: that a component C of the smartspeaker D counts as planned obsolescence; and that any company manufacturing devices with planned obsolescence should not be legally doing business in Norway. The latter obligation would be violated by the fact that M is legally doing business in Norway.

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. **Combined framework dilemma/conflict.** When all arguments are merged in a combined framework, there are multiple extensions;

3. **Integrated framework dilemma/conflict.** When all normative systems are combined to generate an integrated 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.

A moral dilemma or normative conflict is resolved, for example, 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. But uniqueness of extensions is not necessary for moral dilemmas: as long as the extensions do not have conflicting obligations, all previous moral dilemmas can be seen as resolved.

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: 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. This conflict resolution mechanism contains only contextual norms for the preference of some stakeholder over another.

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? Again, we do not permit them. Applying constitutive norms in this context would become a form of institutional wishful thinking.

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

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

$$\mathcal{R}(\mathcal{K}') = \{x \mid a \Rightarrow x \in \mathcal{R}, a \in Cn(\mathcal{K}')\}$$

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

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

- *The obligations are the formulas that can be detached from $I(\mathcal{R}^c, \mathcal{K})$ and $\mathcal{R}^r$. That is, $O(\mathcal{R}^c, \mathcal{R}^r, \mathcal{K}) = \mathcal{R}^r(I(\mathcal{R}^c, \mathcal{K}))$;*

- *The permissions are the formulas that can be detached from $I(\mathcal{R}^c, \mathcal{K})$ and $\mathcal{R}^p$. That is, $P(\mathcal{R}^c, \mathcal{R}^p, \mathcal{K}) = \mathcal{R}^p(I(\mathcal{R}^c, \mathcal{K}))$.*

If we consider only the $\mathcal{R}^c$-norms of a stakeholder s, then we write this set as $\mathcal{R}^c_s$ etc.

Roughly, there is a conflict when I is inconsistent, there is a violation when O is inconsistent with I or with $\mathcal{K}$, 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 $(\mathcal{R}^c, \mathcal{R}^r, \mathcal{R}^p)$ in context $\mathcal{K}$ is a triple (M^c, M^r, M^p) such that:*

- *M^c is a maximal subset of $\mathcal{R}^c$ such that $I(M^c, \mathcal{K})$ is consistent;*
- *M^r and M^p are maximal subsets of $\mathcal{R}^r$ and $\mathcal{R}^p$ such that $I(M^c, \mathcal{K}) \cup O(M^c, M^r, \mathcal{K})$ is consistent, and for all $p \in P(M^c, M^p, \mathcal{K})$, the set $O(M^c, M^r, \mathcal{K}) \cup \{p\}$ is also consistent.*

The norm extension (M^c, M^r, M^p) corresponds to institutional facts $I(M^c, \mathcal{K})$, obligations $O(M^c, M^r, \mathcal{K})$ and permissions $P(M^c, M^p, \mathcal{K})$. A norm extension (M^c, M^r, M^p) is P-maximal if $M^p = \mathcal{R}^p$.

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 deontic options) and moral decision problems as follows.

Definition 4 (Morally sensitive situation)**.** *Given a normative system $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R})$, a morally sensitive situation is defined as any consistent set $mss \subseteq \mathcal{L}$ satisfying: $mss \cap bd[\mathcal{R}] \neq \emptyset$. The set of all possible morally sensitive situations for an ethical agent is denoted as* MSS*.*

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

Definition 5 (Moral decision variable)**.** *Given a normative system $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R})$, the set of all possible moral decision variables of the ethical agent is defined as the $\subseteq$-minimal set $\mathcal{DV}$ that contains $hd[\mathcal{R}^r]$ and is closed under the contrariness function $^-$. The set $\mathcal{DV}$ represents 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)**.** *Given a normative system $\mathcal{N}$, a pair $DP = (mss, DV)$ is a moral decision problem for $\mathcal{N}$, if $mss \in$ MSS is a morally sensitive situation and $DV \subseteq \mathcal{DV}$ is a set of moral decision variables closed under the contrariness function.*

Example 4.1. Let $\mathcal{S} = \{L, H, M, J\}$ be a set of stakeholders, where J is the Jiminy. Let also $Var = \{w_1, w_2, w_3, w_4, i_1, d_1, d_2, d_3, a_1, a_2\}$ be as in Example 3.1. Define the language $\mathcal{L}$ and the contrariness function $^- : \mathcal{L} \mapsto 2^{\mathcal{L}}$ as follows:

$$\mathcal{L} = Var \cup \{\neg v : v \in Var\} \cup \{s \succ s' \mid s \neq s' \text{ for } s, s' \in \mathcal{S}\}$$
$$\bar{a_1} = \{a_2\}, \quad \bar{a_2} = \{a_1, d_2\}, \quad \bar{d_1} = \{d_2, a_2\}, \quad \bar{d_2} = \{a_2, d_1\}, \quad \bar{d_3} = \{d_2\},$$
$$\text{i.e. } a_1 = -a_2 \text{ and } d_1 = -d_2 \text{ and } a_2 = -d_2{}^3$$
$$\overline{s \succ s'} = \{s' \succ s\}, \text{ for all } s \succ s' \in \mathcal{L}.$$

The set $\{w_1, w_2, w_3, w_4\} \subseteq \mathcal{L}$ is a morally sensitive situation. $\mathcal{DV} = \{d_1, d_2, d_3, a_1, a_2\}$ is the set of moral decision variables. Taking $DV = \mathcal{DV}$, we obtain a moral decision problem $DP = (\{w_1, w_2, w_3, w_4\}, \{d_1, d_2, d_3, a_1, a_2\})$.

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 $\mathcal{K}$ of brute facts. Stakeholders are assumed to share the same language $\mathcal{L}$, knowledge base $\mathcal{K}$, and the contrariness function $^-$.

Definition 7 (Stakeholder Argumentation theory)**.** *Let $\mathcal{S} = \{s_1, \ldots, s_n, J\}$ be a set of stakeholders, where J stands for Jiminy. An argumentation theory of a stakeholder $s \in \mathcal{S}$ is a tuple abusively denoted by $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s, \mathcal{K})$ where $(\mathcal{L}, ^-, \mathcal{R}_s)$ is the normative system of s and $\mathcal{K}$ is the set of observations, called the* context*.*

For the normative system of the Jiminy $\mathcal{N}_J = (\mathcal{L}, ^-, \mathcal{R}_J)$, each norm in $\mathcal{R}_J$ will be of the form $\psi_1, \ldots, \psi_k \Rightarrow s_1 \succ s_2$, 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 $\mathcal{S}_0$, i.e., $\mathcal{S}_0 = \mathcal{S} \setminus \{J\}$.

Example 4.2. For each $s \in \mathcal{S} = \{L, H, M, J\}$, the set of norms $\mathcal{R}_s$ is resp. defined by:

$$\mathcal{R}_L = \left\{ \begin{array}{l} w_1 \Rightarrow^r_L d_1, \\ i_1 \Rightarrow^r_L a_1 \end{array} \right\} = \left\{ \begin{array}{c} \textit{D is made by M} \Rightarrow^r_L \textit{M is law Compliant}, \\ \textit{M is business in Norway} \Rightarrow^r_L \textit{to Comply with GDPR} \end{array} \right\}$$
$$\mathcal{R}_H = \left\{ \begin{array}{l} w_2 \Rightarrow^r_H d_2, \\ w_3 \Rightarrow^r_H d_3 \end{array} \right\} = \left\{ \begin{array}{c} \textit{D collects Data} \Rightarrow^r_H \textit{D protects privacy}, \\ \textit{D finds Threat} \Rightarrow^r_H \textit{D reports Threat} \end{array} \right\}$$
$$\mathcal{R}_M = \left\{ \begin{array}{l} w_3 \Rightarrow^r_M a_2, \\ w_4 \Rightarrow^c_M i_1 \end{array} \right\} = \left\{ \begin{array}{c} \textit{D finds Threat} \Rightarrow^r_M \textit{to Collect Data w.o. permission}, \\ \textit{M registered in Norway} \Rightarrow^c_M \textit{M is business in Norway} \end{array} \right\}$$
$$\mathcal{R}_J = \left\{ \begin{array}{l} w_2 \Rightarrow L \succ M, \\ w_3 \Rightarrow L \succ H, \\ \neg w_3 \Rightarrow H \succ L \end{array} \right\} = \left\{ \begin{array}{l} \text{if } w_2, \textit{ } \mathcal{R}_L\textit{-norms take priority over } \mathcal{R}_M\textit{-norms} \\ \text{if } w_3, \textit{ } \mathcal{R}_L\textit{-norms take priority over } \mathcal{R}_H\textit{-norms} \\ \text{if } \neg w_3, \textit{ } \mathcal{R}_H\textit{-norms take priority over } \mathcal{R}_L\textit{-norms} \end{array} \right\}$$

The argumentation theory $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s, \mathcal{K})$ of each stakeholder $s \in \mathcal{S}$ also contains the context:

$$\mathcal{K} = \{w_1, w_2, w_3, w_4\}.$$

Note that $\mathcal{R}_L$, $\mathcal{R}_H$ and $\mathcal{R}_L$ jointly describe the six norms from Example 3.1, while $\mathcal{R}_J$ contains the dilemma-resolving norms of the Jiminy.

In this paper, the notion of argument is defined in terms of the one presented by (Pigozzi & van der Torre, 2018). 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 & van der Torre, 2018), 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.

Definition 8 (Argument)**.** *Let $\mathcal{S} = \{s, \ldots\}$ be a set of stakeholders and $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s, \mathcal{K})$ the argumentation theory of $s \in \mathcal{S}$. An argument A for a conclusion $\mathrm{Conc}(A) = \phi$ is:*

1. **a brute fact argument** $\{\phi\}$ *if* $\phi \in \mathcal{K}$.

2. **an institutional fact argument** $A_1, \ldots, A_n \Rightarrow_s^c \phi$ *if* $A_1, \ldots, A_n$ *are brute or institutional fact arguments and* $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow_s^c \phi$ *is a norm in* $\mathcal{R}_s$.

3. **an obligation argument** $A_1, \ldots, A_n \Rightarrow_s^r \phi$ *if* $A_1, \ldots, A_n$ *are brute or institutional fact arguments such that there exists a norm* $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow_s^r \phi$ *in* $\mathcal{R}_s$.

4. **a permission argument** $A_1, \ldots, A_n \Rightarrow_s^p \phi$ *if* $A_1, \ldots, A_n$ *are brute or institutional fact arguments such that there exists a norm* $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow_s^p \phi$ *in* $\mathcal{R}_s$.

5. **a dilemma resolving argument** $A_1, \ldots, A_n \Rightarrow s_1 \succ s_2$ *if* $A_1, \ldots, A_n$ *are brute or institutional fact arguments such that there exists a norm* $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow s_1 \succ s_2$ *in* $\mathcal{R}_J$. *In this case,* $\mathrm{Conc}(A) = s_1 \succ s_2$. *We indistinctly write such an argument as* $A_1, \ldots, A_n \Rightarrow_J s_1 \succ s_2$.

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

Definition 9 (Argument properties)**.** *For a brute fact argument $\{\phi\}$, we define* $\mathrm{Prem}(\{\phi\}) = \{\phi\}$, $\mathrm{Sub}(\{\phi\}) = \{\phi\}$, $\mathrm{TopNorm}(\{\phi\}) =$ *undefined,* $\mathrm{Norms}(\{\phi\}) = \emptyset$, $\mathrm{Stakeholders}(A) = \emptyset$; *and for an argument $A = A_1, \ldots, A_n \Rightarrow_s^\tau \phi$,*

$$\begin{aligned}
\text{Prem}(A) &= \text{Prem}(A_1) \cup \cdots \cup \text{Prem}(A_n) \\
\text{Sub}(A) &= \text{Sub}(A_1) \cup \ldots \cup \text{Sub}(A_n) \cup \{A\} \\
\text{TopNorm}(A) &= \text{Conc}(A_1), \ldots, \text{Conc}(A_n) \Rightarrow_s^\tau \phi \\
\text{Norms}(A) &= \text{Norms}(A_1) \cup \ldots \cup \text{Norms}(A_n) \cup \{\text{TopNorm}(A)\} \\
\text{Stakeholders}(A) &= \text{Stakeholders}(A_1) \cup \ldots \cup \text{Stakeholders}(A_n) \cup \{s\}
\end{aligned}$$

although in practice we will restrict this set to object level stakeholders, so that $J \notin$ Stakeholders(A). *$Arg(\mathcal{N})$ denotes the sets of all arguments constructed from an argumentation theory* $\mathcal{N} = (\mathcal{L}, \bar{}\,, \mathcal{R}, \mathcal{K})$.

Definition 10 (Institutional facts, obligations and permissions)**.** *Let* $\mathcal{R}^\tau \subseteq \mathcal{R}_{s_1} \cup \cdots \cup \mathcal{R}_{s_n}$, $\tau \in \{r, c, p\}$, *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.*

We also use $OArg(\mathcal{N}) \subseteq Arg(\mathcal{N})$ *to denote the set of obligation arguments in any argumentation framework* $AF = (Arg(\mathcal{N}), Def(\mathcal{N}))$.

Example 4.3. We continue Example 4.2. We may construct a brute fact argument $W_i = \{w_i\}$ for each element $w_i \in \mathcal{K}$. W_i states that w_i is indeed a fact. For each stakeholder $s \in \mathcal{S}$, $W_1, \ldots, W_4$ are arguments in the set $Arg(\mathcal{N}_s)$. Based on these arguments, we may also construct the following:

$A_1 = (W_1 \Rightarrow_L^r d_1)$**:** The manufacturer makes the smart speaker. Hence, it should comply with the law.

$A_2 = (W_2 \Rightarrow_H^r d_2)$**:** The smart speaker is collecting user data. Hence, the privacy of the users should be protected.

$A_3 = (W_3 \Rightarrow_H^r d_3)$**:** The collected (user) information contains a potential critical danger to society. Hence, this information should be reported.

$A_4 = (W_4 \Rightarrow_M^c i_1)$**:** The manufacturer is registered in Norway. This counts as legally doing business in Norway.

$A_5 = (W_3 \Rightarrow_M^r a_2)$**:** The collected (user) information contains a potential critical danger to society. Hence, it ought to collect information without users' explicit consent.

Now we can form $Arg(\mathcal{N}_L) = \{W_1, W_2, W_3, W_4, A_1\}$, $Arg(\mathcal{N}_H) = \{W_1, \ldots, W_4, A_2, A_3\}$ and $Arg(\mathcal{N}_M) = \{W_1, \ldots, W_4, A_4, A_5\}$. Note that no argument exists (yet) that is built upon the norm $i_1 \Rightarrow_L^r a_1$ from Ex. 4.2.

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 (or contraries), meaning that not both of the obligations can be accepted even if both arguments have the same priority. In terms of argumentation theory, we say that these two arguments defeat each other. Meanwhile, when one argument defeats 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., $Op \wedge O\neg p$ 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 the these considerations, the notions of priority relation and defeat relation between arguments are defined as follows.

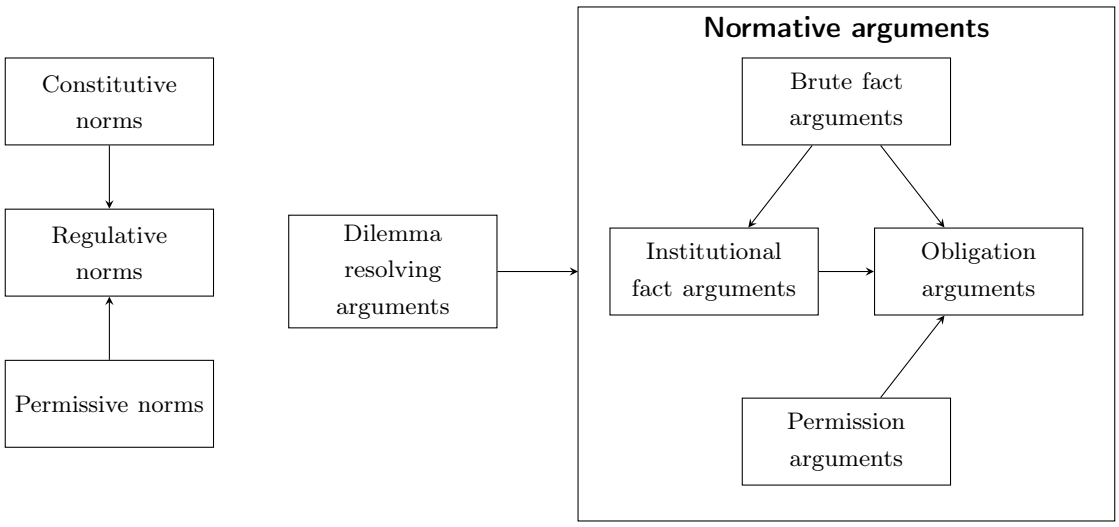


Figure 3: The priority order over different type of norms and arguments. An exiting arrow indicates a higher priority.

Concerning the priority over arguments, according to the normative theory introduced in Section 3, constitutive norms always override regulative norms (otherwise wishful thinking) and so do permissive norms (as they encode exceptions to regulations). So, an institutional fact argument may defeat an obligation argument, a permission argument may defeat an obligation argument, a brute fact argument may defeat an institutional fact argument or an obligation argument, but not vice versa. In addition, brute fact arguments have the highest priority. This is illustrated in Figure 3 and specified in Definition 11. Dilemma resolving arguments will be introduced later on.

Definition 11 (Priority relation between arguments)**.** *Let $\mathcal{A}$ be a set of arguments, e.g. $\mathcal{A} = Arg(\mathcal{N})$ for some $\mathcal{N}$. 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 $A \succeq B$ to denote that A is non-strictly preferred to B and $A \succ B = (A \succeq B$ and $B \not\succeq A)$ to denote that A is preferred to B. We have:

- *For $A, B \in \mathcal{A}^\tau$, $A \succeq B$ any $\tau \in \{b, c, r\}$.*
- *For $A \in \mathcal{A}^b$ and $B \in \mathcal{A}^c \cup \mathcal{A}^r$, $A \succ B$.* [4]
- *For $A \in \mathcal{A}^c$ and $B \in \mathcal{A}^r$, $A \succ B$.*
- *For $A \in \mathcal{A}^p$ and $B \in \mathcal{A}^r$, $A \succ B$.*

For a fixed type $\tau \neq p$, any two arguments $A, B \in \mathcal{A}^\tau$ are equally preferred: $A \succeq B \succeq A$. (And for $\tau = p$, incomparable: $A \not\succeq B \not\succeq A$). If desired, a strict preference can be enforced between them, e.g. $A \succ' B$, in a more refined priority relation $\succ' \supset \succ$. 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 dilemmas, 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.[5]

Definition 12 (Attacks and defeats)**.** *Let $\mathcal{A}$ be a set of arguments. For any $A, B \in \mathcal{A}$, A attacks B, iff $\mathrm{Conc}(A) \in \bar{\phi}$ for some $B' \in \mathrm{Sub}(B)$ and $\mathrm{Conc}(B') = \phi$. In this case, we say that A attacks B at B'. We say that A defeats B iff:*

- *A attacks B at B' and $A \succeq B'$* *(direct defeat)*
- *or B extends some $B' \in \mathrm{sub}(B)$ that attacks A at A and $B' \prec A$.* *(reverse defeat)*

The set of defeats over the arguments $Arg(\mathcal{N})$ from an argumentation theory $\mathcal{N}$ is denoted $Def(\mathcal{N})$.

Definition 13 (Argumentation frameworks: individual, combined)**.** *Given a set of stakeholders $\mathcal{S}_0 = \{s, \ldots, \}$, we call:*

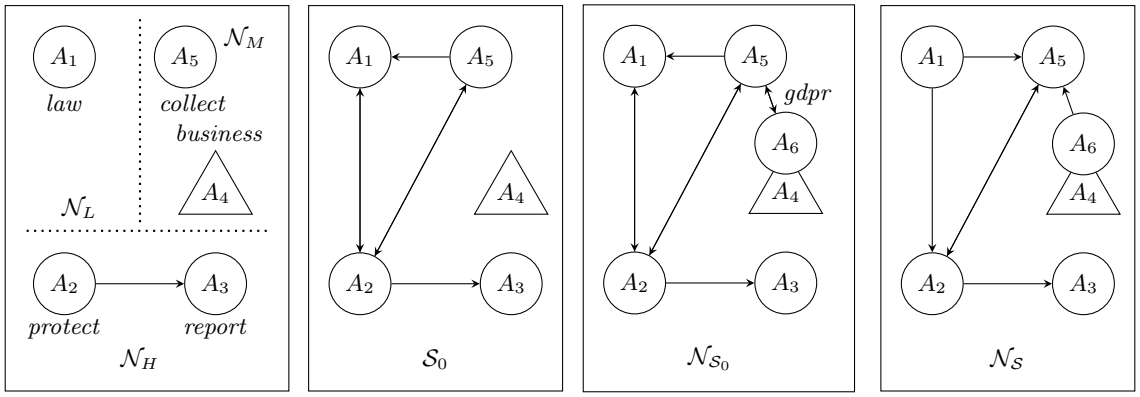
$AF(\mathcal{N}_s) = \quad (Arg(\mathcal{N}_s), Def(\mathcal{N}_s))$ *an individual argumentation framework of s.*

Gathering all arguments from stakeholders defines the set $Arg(\mathcal{S}_0) = \bigcup_{s \in \mathcal{S}_0} Arg(\mathcal{N}_s)$. By letting $Def(\mathcal{S}_0)$ be the defeat relation over this set, we obtain:

$AF(\mathcal{S}_0) = \quad (Arg(\mathcal{S}_0), Def(\mathcal{S}_0))$ *a combined argumentation framework.*

Example 4.4. Continue Example 4.3. Three individual argumentation frameworks of the stakeholders L, H and M are illustrated in Figure 4a. In $Arg(\mathcal{N}_H)$ we see that argument A_2 attacks argument A_3, because reporting the information collected from users is in conflict with protecting their privacy.

4. A condition one might want to impose upon the language $\mathcal{L}$ is that brute facts are disjoint from institutional facts —so that priorities of type $\mathcal{A}^b \succ \mathcal{A}^c$ would not be needed. In practice this condition cannot always be enforced, see the discussion in (Pigozzi & van der Torre, 2018, Sec. 4.3).
5. Attacks represent logical conflicts based on the contrariness relation $^-$. While conceptually there is no real conflict between a permission for p and a fact or permission for $\neg p$ (i.e. only with an obligation $\neg p$) we keep the definition of attack simple and manage this lack of real conflict via incomparability under the priority relation $\succeq$.



(a) Individual frameworks (b) Combined (c) Integrated (d) Reduced

Figure 4: Argumentation frameworks at 4 levels. Context arguments $W_1, \ldots, W_4$ are omitted. Obligation and institutional arguments are represented as circles resp. triangles (next to their conclusions), and defeats as arrows. A subargument of an argument is depicted as partly behind it. (a) contains dilemmas between the conclusions of $\{A_2, A_3\}$, of $\{A_1, A_2\}$, and of $\{A_1, A_5\}$. (b) makes all dilemmas explicit as defeats. (c) brings upon a new dilemma involving $\{A_5, A_6\}$. (d) revises the defeat relation based on Jiminiy's arguments for preferences.

$(A_2 \to A_3)$ Since $d_2 \in \overline{d_3}$ and $A_2 \succeq A_3$, this conflict is also a defeat: $Def(\mathcal{N}_H) = \{(A_2, A_3)\}$.

(since $d_3 \notin \overline{d_2}$, for the defeat $A_3 \to A_2$ we would need the stronger preference $A_3 \succ A_2$).

The combined argumentation framework $AF(\mathcal{S}_0)$ in Figure 4b adds the following defeats:

$(A_1 \leftrightarrow A_2)$ the conflict between *M is law Compliant* and *D protects privacy* is a mutual defeat: $d_1 = -d_2$ and $A_1 \preceq\succeq A_2$;

$(A_5 \to A_1)$ *to Collect Data w.o. permission* is in conflict with *M is law Compliant*: $a_2 \in \overline{d_1}$ and $A_5 \preceq\succeq A_1$;

$(A_5 \leftrightarrow A_2)$ *to Collect Data w.o. permission* is in mutual conflict with *D protects privacy*: $a_2 = -d_2$ and $A_5 \preceq\succeq A_2$.

In summary, $Def(\mathcal{S}_0) = \{(A_2, A_3), (A_1, A_2), (A_2, A_1), (A_5, A_1), (A_5, A_2)\}$.

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 $\mathcal{E} \subseteq Arg$ is *conflict-free* if and only if there exist no arguments A and B in $\mathcal{E}$ such that $(A, B) \in Def$. Argument $A \in Arg$ is *defended* by a set $\mathcal{E} \subseteq Arg$ (also called A is *acceptable* with respect to $\mathcal{E}$) if and only if for all $B \in Arg$, if $(B, A) \in Def$, then there exists $C \in \mathcal{E}$ 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 14 (Argumentation semantics)**.** *Let $AF = (Arg, Def)$ be an argumentation framework, and $\mathcal{E} \subseteq Arg$ a set of arguments.*

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

We use $\sigma \in \{co, pr, gr, st\}$ to indicate the complete, preferred, grounded, and stable semantics.

Recall that in these semantics $\sigma \in \{co, pr, gr, st\}$, their extensions are complete: $\sigma(AF) \subseteq co(AF)$; and hence also admissible: $\mathcal{E} \in \sigma(AF)$ implies that $\mathcal{E}$ defends all of $\mathcal{E}$. We focus first on a combination of argument types whose logical conflicts do matter, e.g. the factual, institutional and obligation arguments. For any extension $\mathcal{E}$, we call the set $\mathcal{E}^{bcr} = \mathcal{E} \setminus \mathcal{A}^p$ the *bcr*-fragment of $\mathcal{E}$.

Lemma 4.5 (Rationality postulates)**.** *The rationality postulates (Caminada & Amgoud, 2007) hold for the bcr-fragment $\mathcal{E}^{bcr}$ of any σ-extension $\mathcal{E}$ under $\sigma \in \{co, pr, gr, st\}$. In particular, direct consistency is satisfied by any bcr-fragment $\mathcal{E}^{bcr}$ of some $\mathcal{E} \in \sigma(AF)$; subargument closure holds for any extension $\mathcal{E} \in \sigma(AF)$ as well.*

A similar direct consistency result can be shown for all obligation-permission pairs.

Fact 4.6. *For any σ-extension $\mathcal{E}$ with $\sigma \in \{co, pr, gr, st\}$, the set $\mathrm{Conc}((\mathcal{E} \cap \mathcal{A}^r) \cup \{B\})$ is consistent, where B is an arbitrary permission argument in the extension, i.e. $B \in \mathcal{E} \cap \mathcal{A}^p$.*

The correspondence between (the outputs of) semantic extensions and norm extensions is at best partial, in the sense of each complete extension *being contained in* a norm extension. Only the outputs of stable extensions match those of a norm extension.

Proposition 4.7. *Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory with $\mathcal{R} = \mathcal{R}^c \cup \mathcal{R}^r \cup \mathcal{R}^p$. (1) For any extension $\mathcal{E} \in \sigma(AF(\mathcal{N}))$ under $\sigma \in \{co, pr, gr, st\}$ there exists (M^c, M^r, M^p), a norm extension of $(\mathcal{R}^c, \mathcal{R}^r, \mathcal{R}^p)$ in context $\mathcal{K}$, such that:*

(i) $I(M^c, \mathcal{K}) \supseteq \mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$;

(ii) $O(M^c, M^r, \mathcal{K}) \supseteq \mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^r)$;

(iii) $P(M^c, M^p, \mathcal{K}) \supseteq \mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^p)$.

(2) For the stable case $\sigma = st$, the inclusions in (i)–(iii) are in fact identities: (i′) $I(M^c, \mathcal{K}) = \mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$; (ii′) $O(M^c, M^r, \mathcal{K}) = \mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^r)$; (iii′) $P(M^c, M^p, \mathcal{K}) = \mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^p)$.

Conversely to Prop. 4.7(1), certain conditions verify that all P-maximal norm extensions M contain (the rules of) some extension $\mathcal{E}$ in those semantics $\sigma \in \{co, pr, gr\}$ that grant the existence of extensions. A condition verifying this claim is the symmetry of contraries: $\phi \in \overline{\psi}$ iff $\psi \in \overline{\phi}$.

Proposition 4.8. *Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory with a symmetric contrariness function $^-$. Then, any P-maximal norm extension $M = (M^c, M^r, \mathcal{R}^p)$ in $\mathcal{K}$ contains a σ-extension $\mathcal{E}$ in the sense of the (i)–(iii) from Prop. 4.7 for any $\sigma \in \{co, pr, gr\}$.*

Definition 15 (Naive semantics)**.** *An argumentation semantics not based on the idea of admissibility is the naive semantics $\sigma = na$, defined as:*

- *$\mathcal{E}$ is a naive extension, denoted $\mathcal{E} \in na(AF)$, iff the set $\mathcal{E}$ is $\subseteq$-maximally conflict free.*

The naive semantics provides a better correspondence with norm extensions, as both notions are defined by maximal conflict freeness and resp. consistency. In fact, a 1-1 correspondence exists between naive- and norm-extensions at the level of triggered rules.

Proposition 4.9. *Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory, inducing the argumentation framework $AF = (Arg(\mathcal{N}), Def(\mathcal{N}))$. (1) For any naive extension $\mathcal{E} \in na(AF)$, it holds that $\mathcal{E} = Arg(\mathcal{N}_M)$ for some norm extension M in $\mathcal{K}$. (2) For any P-maximal norm extension M in $\mathcal{K}$, the set $\mathcal{E}_M = Arg(\mathcal{N}_M)$ is a naive extension: $\mathcal{E}_M \in na(AF)$.*

Although all kinds of defeats have an impact on the evaluation of argumentation frameworks, the proposed system will particularly seek (and address) moral dilemmas.

Given an argumentation framework $AF(\mathcal{N}) = (Arg(\mathcal{N}), Def(\mathcal{N}))$ of an argumentation theory $\mathcal{N}$, we denote the set of obligation arguments by $\mathrm{OArg}(\mathcal{N})$. Given an extension $\mathcal{E} \in \sigma(AF(\mathcal{N}))$ we also let $Obl(\mathcal{E}) = \{\mathrm{Conc}(A) \mid A \in \mathcal{E} \cap \mathrm{OArg}(\mathcal{N})\}$ be the set of obligations in the conclusions of $\mathcal{E}$.

Definition 16 (Moral dilemma)**.** *Let $\mathcal{C}$ be a collection of argument extensions for some decision problem $DP = (mss, DV)$. We say that $\mathcal{C}$ contains a moral dilemma if a pair of obligation arguments A, B exist in some extensions of $\mathcal{C}$, say $A \in \mathcal{E}_1 \in \mathcal{C}$ and $B \in \mathcal{E}_2 \in \mathcal{C}$, such that $\mathrm{Conc}(A) \in \overline{\mathrm{Conc}(B)}$ and these two contrary obligations are in DV. In other words, there exist $\mathcal{E}_1, \mathcal{E}_2 \in \mathcal{C}$ such that*

$(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV$ is inconsistent with respect to the contrariness function $^-$.

Given the argumentation theory $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s, \mathcal{K})$ of each stakeholder $s \in S$, and a decision problem, we distinguish four ways (i.e. four collections $\mathcal{C}$ of extensions) to check whether there is a dilemma and, if needed, use the dilemma resolving norms to deal with the dilemma:

1. First, we consider the normative system of each object level stakeholder independently. In this case, we compute the extensions of the corresponding argumentation frameworks and check whether there is a dilemma between the extensions of one or more object level stakeholders.

2. Second, we consider the arguments of all stakeholders together. In this case, we construct a single argumentation framework to check whether there is a dilemma. Each argument still consists of a set of norms from the normative system of a single object level stakeholder.

3. Third, we put all normative systems together, and a unified argumentation theory to check whether there is a dilemma. Arguments now combine norms from different stakeholders.

4. Fourth, we use the Jiminy to decide among the stakeholders the most competent for the dilemma.[6]

See Section 6 for an explanation of these four levels and their role in different application domains. At any of these four levels, the Jiminy submits as its moral recommmendation in case no dilemma is found, namely as the set of obligations occurrying in at least one of the semantic extensions.

Definition 17 (Moral recommendation at i-th level)**.** *If no dilemma exists in the set of extensions $\{\mathcal{E}_1, \ldots, \mathcal{E}_k\}$ at level i, then the moral recommendation or output at i is: $Obl(\mathcal{E}_1) \cup \ldots \cup Obl(\mathcal{E}_k)$.*

Despite obligation arguments having the lowest priorities among arguments, the checking of dilemmas makes the system credulous about obligations (it accepts an obligation if it belongs to at least an extension) and skeptical about institutional facts and permissions.

1st level dilemmas: Individual Frameworks. Let us fix an argumentation theory $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s, \mathcal{K})$ for each stakeholder $s \in \mathcal{S}_0$.

Definition 18 (Individual Frameworks $\mathcal{IF}$)**.** *The set of* individual frameworks *is:*

$$\mathcal{IF} = \{AF(\mathcal{N}_s) : s \in \mathcal{S}_0\}$$

where $AF(\mathcal{N}_s) = (Arg(\mathcal{N}_s), Def(\mathcal{N}_s))$ is the argumentation framework of stakeholder $s \in \mathcal{S}_0$.

For a reference, let us define a zero level dilemma as any moral dilemma between a pair of (unfiltered) arguments of stakeholders $\mathcal{C}_0 = \{Arg(\mathcal{N}_{s_0}), \ldots, Arg(\mathcal{N}_{s_0})\}$. (The four levels of dilemma checking make use of argumentation semantics to filter out some of these arguments and dilemmas.) Let us stress that dilemmas depend on the contrariness relation rather than the defeat relation(s).

Definition 19 ($\mathcal{IF}$ dilemma checking and resolving)**.** *Let $DP = (mss, DV)$ be a decision problem and σ an argumentation semantics: $\sigma \in \{co, gr, pr, st\}$. A first level (or $\mathcal{IF}$) dilemma with respect to DP under σ is any moral dilemma in the collection $\mathcal{C}_1 = \sigma(AF(\mathcal{N}_{s_0})) \cup \ldots \cup \sigma(AF(\mathcal{N}_{s_n}))$. That is, an $\mathcal{IF}$ dilemma exists if there are $\mathcal{E}_1 \in \sigma(AF), \mathcal{E}_2 \in \sigma(AF')$ for some $AF, AF' \in \mathcal{IF}$ such that*

- *$(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV$ is inconsistent with respect to the contrariness function $^-$.*

6. The source of the Jiminy priorities is domain specific. We assume that the set of norms of Jiminy is given.

Otherwise, if for all $\mathcal{E} \in \sigma(AF), \mathcal{E}' \in \sigma(AF')$ *it holds that* $(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV = \emptyset$*, then there is no* $\mathcal{IF}$ *dilemma and all zero level dilemmas have been resolved at the first level by* σ*.*

Example 4.10. Continue Example 4.4. The argumentation frameworks of stakeholders L, H and M (illustrated in Figure 5a) have the preferred extensions $\mathcal{E}_1 = \{W_1, W_2, W_3, W_4, A_1\}$, $\mathcal{E}_2 = \{W_1, \ldots, W_4, A_2\}$ and $\mathcal{E}_3 = \{W_1, \ldots, W_4, A_4, A_5\}$ respectively. $DP = (mss, DV)$ was defined in Example 4.1 by $mss = \{w_1, w_2, w_3, w_4\} = \mathcal{K}$ and $DV = \{d_1, d_2, d_3, a_1, a_2\}$. From $Obl(\mathcal{E}_1) = \{d_1\}$, $Obl(\mathcal{E}_2) = \{d_2\}$ and $Obl(\mathcal{E}_3) = \{a_2\}$ a first level dilemma exists between each pair of extensions:

$$\{\mathcal{E}_1, \mathcal{E}_2\}: \quad d_1 = -d_2; \qquad \{\mathcal{E}_3, \mathcal{E}_1\}: \quad a_2 \in \overline{d_1} \qquad \{\mathcal{E}_2, \mathcal{E}_3\}: \quad a_2 = -d_2.$$

2nd level dilemmas: Combined Framework. For the second level checking of dilemmas, we check the combined argumentation framework $AF(\mathcal{S}_0)$ (Def. 13), consisting of the arguments of all stakeholders $Arg(\mathcal{S}_0) = \bigcup_{s \in \mathcal{S}_0} Arg(\mathcal{N}_s)$ and the defeat relation $Def(\mathcal{S}_0)$ induced by them.

Fact 4.11. *Given a set of individual argumentation frameworks* $\mathcal{IF} = \{AF(\mathcal{N}_s) \mid s \in \mathcal{S}_0\}$ *and the combined argumentation framework* $AF(\mathcal{S}_0) = (Arg(S_0), Def(\mathcal{S}_0))$ *at the second level, it holds that* $Def(\mathcal{S}_0) \supseteq \bigcup_{s \in \mathcal{S}_0} Def(\mathcal{N}_s)$*.*

Proof. If A defeats B according to $(A, B) \in Def(\mathcal{N}_s)$ for all $s \in \mathcal{S}_0$, A still defeats B when A and B are in $Arg_{\mathcal{S}_0}$ according to Definition 12. According to Definition 13, $(A, B) \in Def(\mathcal{S}_0)$. So, it holds that $Def(\mathcal{S}_0) \supseteq \bigcup_{s \in \mathcal{S}_0} Def(\mathcal{N}_s)$. □

Definition 20 (Combined framework dilemma checking and resolving)**.** *Let* $DP = (mss, DV)$ *be a decision problem, and* $AF(\mathcal{S}_0) = (Arg(\mathcal{S}_0), Def(\mathcal{S}_0))$ *the combined framework. A second level (or* $AF(\mathcal{S}_0)$*) dilemma with respect to* DP *under* σ *is any moral dilemma in* $\mathcal{C}_2 = \sigma(AF(\mathcal{S}_0))$*. That is, such a dilemma exists if there are* $\mathcal{E}_1, \mathcal{E}_2 \in \sigma(AF(\mathcal{S}_0))$ *such that*

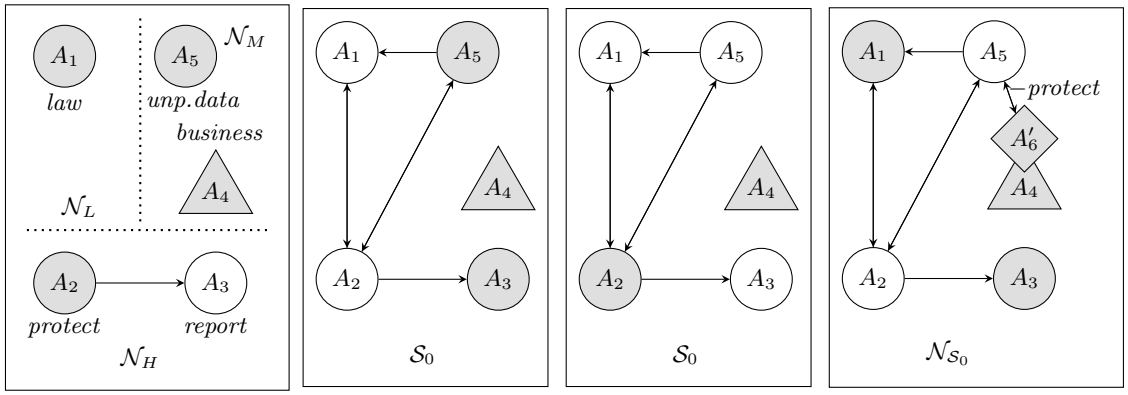
$$(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV \text{ is inconsistent with respect to } {}^{-}.$$

Otherwise, if for all $\mathcal{E}_1, \mathcal{E}_2 \in \sigma(AF_{\mathcal{S}_0})$*,* $(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV = \emptyset$*, then there is no second level dilemma and all first level dilemmas are resolved at the second level by* σ*.*

Example 4.12. Continue Example 4.3. By combining the stakeholders' arguments in AF_L, AF_H and AF_M in Figure 4.4a, we get a combined argumentation framework $AF(\mathcal{S}_0)$ illustrated in Figure 4.4b. Figures 5b–5c illustrate the two preferred extensions: $\mathcal{E}_1 = \{W_1, \ldots, W_4, A_5, A_3, A_4\}$ and $\mathcal{E}_2 = \{W_1, \ldots, W_4, A_2, A_4\}$, each giving the obligations $Obl(\mathcal{E}_1) = \{a_2, d_3\}$ and $Obl(\mathcal{E}_2) = \{d_2\}$. The dilemmas, now between $\mathcal{E}_1$ and $\mathcal{E}_2$, update as follows from the first to the second level:

solved at 2nd level	*persist from 1st level*	*new in 2nd level*
$d_1 = -d_2$ $a_2 \in \overline{d_1}$	$\{A_5, A_2\}: \ a_2 = -d_2$	$\{A_2, A_3\}: \ d_2 \in \overline{d_3}$

At this level, the smart device cannot decide between {*collecting information without permission*, *reporting the potential threat* } and { *protecting users' privacy* }.



(a) Individual frameworks (b) Combined (c) Combined (d) Integrated

Figure 5: (a)–(c) Preferred extensions for the individual and combined argumentation frameworks in the running example (Ex. 4.10–4.12) and the alternative example (Ex. 4.15). The arguments in gray belong to a preferred extension in its AF. (d) The preferred extension for the alternative Example 4.15 contains a permission argument A_6'. Since this extension is unique all second level dilemmas are solved at the third level.

3rd level dilemmas: Integrated Framework. 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 21 (Integrated argumentation theory, frameworks)**.** *Let $\mathcal{S}_0 = \{s_1, \ldots, s_n\}$ be the set of object-level of stakeholders s, each s endowed with a normative system $(\mathcal{L}, ^-, \mathcal{R}_s)$. Given a context $\mathcal{K}$,*

$$\mathcal{N}_{\mathcal{S}_0} = \quad (\mathcal{L}, ^-, \mathcal{R}_{\mathcal{S}_0}, \mathcal{K}) \text{ with } \mathcal{R}_{\mathcal{S}_0} = \bigcup_{s \in \mathcal{S}_0} \mathcal{R}_s \quad \text{is an integrated argumentation theory.}$$

Such $\mathcal{N}_{\mathcal{S}_0}$ gives rise to an integrated framework: $AF(\mathcal{N}_{\mathcal{S}_0}) = (Arg(\mathcal{N}_{\mathcal{S}_0}), Def(\mathcal{N}_{\mathcal{S}_0}))$.

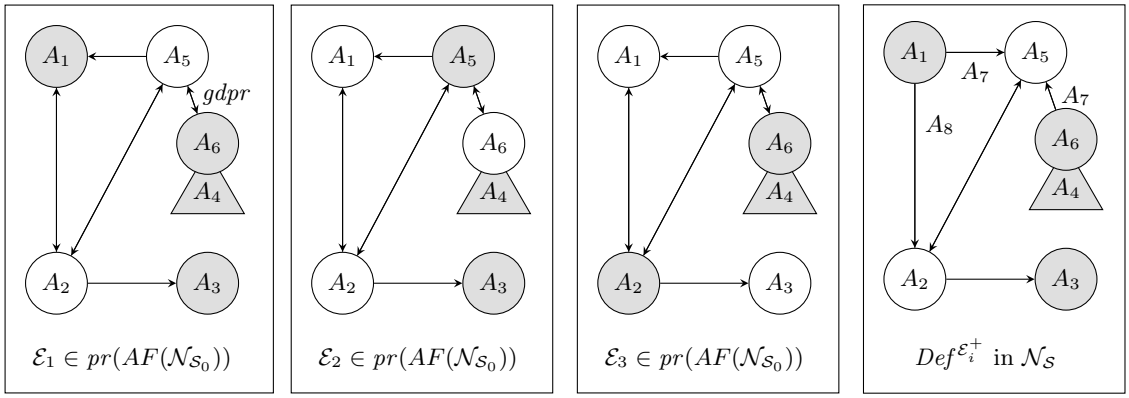
Definition 22 (Integrated framework dilemma checking and resolving)**.** *Let $DP = (mss, DV)$ be a decision problem, σ a semantics and $\mathcal{N}_{\mathcal{S}_0}$ an integrated framework. A third level (or $\mathcal{N}_{\mathcal{S}_0}$) dilemma for DP under σ is any moral dilemma in $\mathcal{C}_3 = \sigma(AF(\mathcal{N}_{\mathcal{S}_0}))$. That is, a dilemma exists if there are $\mathcal{E}_1, \mathcal{E}_2 \in \sigma(AF(\mathcal{N}_{\mathcal{S}_0}))$ such that*

$$(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV \text{ is inconsistent with respect to } ^-.$$

Otherwise, if for all $\mathcal{E}_1, \mathcal{E}_2 \in \sigma(AF(\mathcal{N}_{\mathcal{S}_0}))$, $(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV = \emptyset$, then there is no dilemma at the third level and all second level dilemmas are resolved at the third level.

Fact 4.13. *Given a combined argumentation framework $AF(\mathcal{S}_0) = (Arg(\mathcal{S}_0), Def(\mathcal{S}_0))$ at the second level and an integrated argumentation framework $AF(\mathcal{N}_{\mathcal{S}_0}) = (Arg(\mathcal{N}_{\mathcal{S}_0}), Def(\mathcal{N}_{\mathcal{S}_0}))$ at the third level, it holds that $Arg(\mathcal{S}_0) \subseteq Arg(\mathcal{N}_{\mathcal{S}_0})$ and $Def(\mathcal{S}_0) \subseteq Def(\mathcal{N}_{\mathcal{S}_0})$.*

Proof. According to Definition 9, for all $A \in Arg(\mathcal{S}_0)$ in the combined argumentation framework, A can also be constructed from the integrated argumentation theory $\mathcal{N}_{\mathcal{S}_0}$ (using the rules of one stakeholder), and therefore $A \in Arg(\mathcal{N}_{\mathcal{S}_0})$. So, $Arg(\mathcal{S}_0) \subseteq Arg(\mathcal{N}_{\mathcal{S}_0})$. On the



(a) Integrated (b) Integrated (c) Integrated (d) Reduced

Figure 6: The running example (Ex. 4.14–4.16). (a)–(c) The three extensions of the integrated framework. (d) The dilemma resolving arguments A_7, A_8 revise the defeat relation: the priority between A_1 and A_5 is reversed, and the priorities between $\{A_1, A_2\}$ and $\{A_5, A_6\}$ become asymmetric. The extension $\mathcal{E}_1^+ = \mathcal{E}_1 \cup \{A_7, A_8\}$ of $AF(\mathcal{N}_{\mathcal{S}})$ is compatible with the revised priority $\succeq^{\mathcal{E}_1^+}$.

other hand, for all $(A, B) \in Def(\mathcal{S}_0)$, A and B are in $Arg(\mathcal{S}_0)$ and therefore in $Arg(\mathcal{N}_{\mathcal{S}_0})$. According to definition 12, any element $(A, B) \in Def(\mathcal{S}_0)$ with $A, B \in Arg(\mathcal{S}_0)$ is defined from the internal structure of A, B, the contrariness function $^-$ and the preference relation $\succeq$, which do not change when A and B are considered in $Arg(\mathcal{N}_{\mathcal{S}_0})$. Therefore, we also have $(A, B) \in Def(\mathcal{N}_{\mathcal{S}_0})$. So, $Def(\mathcal{S}_0) \subseteq Def(\mathcal{N}_{\mathcal{S}_0})$. □

Example 4.14. Using rules from both $\mathcal{R}_M$ and $\mathcal{R}_L$, the integrated argumentation framework $AF(\mathcal{N}_{\mathcal{S}_0})$ generates a new argument (not present in the combined $AF(\mathcal{S}_0)$), namely

$$A_6 = A_4 \Rightarrow^r a_1 \quad \text{with conclusion } a_1 = \textit{to Comply with GDPR}$$

The preferred extensions are: $\mathcal{E}_1 = \{W_1, \ldots, W_4, A_1, A_3, A_4, A_6\}$ and $\mathcal{E}_2 = \{W_1, \ldots, W_4, A_5, A_4, A_3\}$ and $\mathcal{E}_3 = \{W_1, \ldots, W_4, A_2, A_4, A_6\}$. They give rise to the obligations $Obl(\mathcal{E}_1) = \{d_1, d_3, a_1\}$ and $Obl(\mathcal{E}_2) = \{a_2, d_3\}$ and $Obl(\mathcal{E}_3) = \{d_2, a_1\}$. The third level dilemmas are:

reinstated from 1st level	*persist from 2nd level*	*new in 3rd level*
$\{A_1, A_2\}: \ d_1 = -d_2$	$\{A_5, A_2\}: \ a_2 = -d_2$	$\{A_6, A_5\}: \ a_1 = -a_2$
$\{A_5, A_1\}: \ a_2 \in \overline{d_1}$	$\{A_2, A_3\}: \ d_2 \in \overline{d}_3$	

At this level, the smart speaker cannot decide between the following three sets of obligations: $\{law, report, gdpr\}$ and $\{collect, report\}$ and finally $\{gdpr, protect\}$.

In the running example, the integrated framework actually worsens the situation by adding new dilemmas to the old ones, and resolving none of them. For the next scenario, in contrast, all dilemmas at the second level are resolved at the third level.

Example 4.15 (Alternative example)**.** Replace in Example 4.2 the GDPR norm (from stakeholder L) with a permission to not protect users' privacy. That is, replace $\mathcal{R}_L =$

$\{w_1 \Rightarrow^r_L d_1, i_1 \Rightarrow^r_L a_1\}$ with $\mathcal{R}'_L = \{w_1 \Rightarrow^r_L d_1, i_1 \Rightarrow^p_L -d_2\}$. The resulting combined and individual frameworks $AF(\mathcal{S}_0)$ and each $AF(\mathcal{N}_s)$ consist of the same defeats, extensions and dilemmas as in Ex. 4.10–4.12. For the combined framework $AF(\mathcal{S}_0)$, the two extensions (Figure 5b–5c) give two second level dilemmas:

$$a_2 = -d_2 \qquad d_2 \in \overline{d_3}.$$

The integrated framework $AF(\mathcal{N}_{\mathcal{S}_0})$ is now defined from $\mathcal{N}_{\mathcal{S}_0} = (\mathcal{L}, ^-, \mathcal{R}'_L \cup \mathcal{R}_M \cup \mathcal{R}_H, \mathcal{K})$. Its set of arguments replaces A_6 with $A'_6 = A_4 \Rightarrow^p -d_2$. This argument defeats both A_2 and A_5. As a result, there is only one preferred extension $\mathcal{E} = \{W_1, \ldots, W_4, A_1, A_3, A_4, A'_6\}$. Hence, the third level contains no dilemmas and resolves all second level dilemmas —see Figure 5d. Under this set of norms, the smart device decides to fulfil $Obl(\mathcal{E}) = \{$ *comply with the law*, *report potential threat* $\}$.

4th level dilemmas: Reduced Framework. Once we combine the stakeholders' norms with the dilemma resolving norms from the Jiminy, we generate all the dilemma resolving arguments. Recall that the Jiminy arguments attack each other in case they have contrary conclusions: $s' \succ s$ and $s \succ s'$. Conclusions of this form, if taken from a conflict-free set of arguments $\mathcal{E}$, induce a new priority among arguments $\succeq \;\longmapsto\; \succeq^{\mathcal{E}}$ and, as a result, a revision of the defeat relation $Def \;\longmapsto\; Def^{\mathcal{E}}$.

Definition 23 (Reduced argumentation framework)**.** *Let $AF = (Arg, Def)$ be an argumentation framework. A conflict-free set $\mathcal{E} \subseteq Arg$ induces the following preference relation $R^{\mathcal{E}} \subseteq Arg \times Arg$:*

$(A, B) \in R^{\mathcal{E}}$ *iff* $\text{Stakeholder}(A) \setminus \text{Stakeholder}(B) \neq \emptyset$, *and for all* $s_A \in \text{Stakeholder}(A) \setminus \text{Stakeholder}(B)$ *and all* $s_B \in \text{Stakeholder}(B) \setminus \text{Stakeholder}(A)$, $s_A \succ s_B \in \text{Conc}(\mathcal{E}^J)$.

The revision of the priority $\succeq$ (Def. 11) by such $R^{\mathcal{E}}$, denoted $\succeq^{\mathcal{E}}$, is defined by:

$$\succeq^{\mathcal{E}} = \; (\succeq \setminus (R^{\mathcal{E}})^{-1}) \cup R^{\mathcal{E}}.$$

A reduced argumentation framework with respect to $\mathcal{E}$ is a pair $AF^{\mathcal{E}} = (Arg, Def^{\mathcal{E}})$, where $Def^{\mathcal{E}}$ is the defeat relation (Def. 12) induced by the revised priority $\succeq^{\mathcal{E}}$.

Example 4.16. Continue Example 4.14. After adding norms from the Jiminy to the integrated argumentation theory, we may construct the argumentation framework $AF(\mathcal{N}_{\mathcal{S}}) = (Arg, Def)$. The set $Arg = \{W_1, \ldots, A_6, A_7, A_8\}$ expands the arguments of integrated framework with:

$$A_7 = W_2 \Rightarrow L \succ M \qquad \text{and} \qquad A_8 = W_3 \Rightarrow L \succ H.$$

If one expands the preferred extension $\mathcal{E}_1$ of the integrated framework (Fig. 6(a)) with the new arguments, say $\mathcal{E}_1^+ = \mathcal{E}_1 \cup \{A_7, A_8\}$, one obtains $R^{\mathcal{E}_1^+} = \{(A_1, A_2), (A_1, A_5), (A_6, A_5)\}$. The revised defeat $Def^{\mathcal{E}_1^+}$ shown in Fig. 6(d) reverses or disables some of the original defeats in Def.

A reduced argumentation framework $AF^{\mathcal{E}}$ depends on which arguments for priorities in $\mathcal{E}$ are selected in the original framework AF. For this reason, we use a two-stage approach to

obtain the extensions of $AF(\mathcal{N}_\mathcal{S})$, based on the approach introduced by Brewka (1994). First, we compute conflict-free sets $\mathcal{E}$ of arguments without considering the priority information contained in the dilemma resolving arguments of $\mathcal{E}$; after considering this priority information, each set $\mathcal{E}$ determines a new priority $\succeq^\mathcal{E}$ and defeat relation $Def^\mathcal{E}$. Then we check the compatibility of each set $\mathcal{E}$ with the priority $\succeq^\mathcal{E}$ induced by it. Formally, we say that $\mathcal{E}$ is compatible with respect to the priority relation $\succeq^\mathcal{E}$ of its dilemma resolving arguments if and only if $\mathcal{E}$ is an extension under the new defeat $Def^\mathcal{E}$. In the terms of Brewka (1994), such a set $\mathcal{E}$ will survive if it can be reconstructed after the priority information from its dilemma resolving arguments is considered.[7]

Definition 24 (Compatibility)**.** *We say that the priority $\succeq^\mathcal{E}$ contained in $\mathcal{E}$ is compatible with $\mathcal{E}$ if and only if $\mathcal{E} \in \sigma(AF^\mathcal{E})$.*

Example 4.17. Continue Example 4.16. Expanding the sets in Fig. 6(a)–(b), namely $\mathcal{E}_1^+ = \mathcal{E}_1 \cup \{A_7, A_8\}$ and $\mathcal{E}_2^+ = \mathcal{E}_2 \cup \{A_7, A_8\}$, leads to the same defeat $Def^{\mathcal{E}_1^+} = Def^{\mathcal{E}_2^+}$ and reduced framework $AF^{\mathcal{E}_1^+} = AF^{\mathcal{E}_2^+}$. This framework has one preferred extension $\mathcal{E}_1^+$ (Fig. 6(d)). Hence,

- $\mathcal{E}_1^+ \in pr(AF^{\mathcal{E}_1^+})$, and so the priority $\succeq^{\mathcal{E}_1^+}$ it contains is compatible with $\mathcal{E}_1^+$.
- $\mathcal{E}_2^+ \notin pr(AF^{\mathcal{E}_2^+})$, and so the priority $\succeq^{\mathcal{E}_2^+}$ it contains is not compatible with $\mathcal{E}_2^+$. (The same applies to the expansion $\mathcal{E}_3^+ = \mathcal{E}_3 \cup \{A_7, A_8\}$ of the thrid extension in Fig. 6(c).)

It only remains to define how we find the Jiminy's recommendations in the original framework $AF(\mathcal{N}_\mathcal{S})$. To this end, we focus on the constitutive and dilemma resolving norms, i.e. the norms that determine the new priority and defeat.

Definition 25 (Jiminy argumentation theory, framework)**.** *Let $\mathcal{S} = \mathcal{S}_0 \cup \{J\}$ be the set of all stakeholders s including J = Jiminy, each with a normative system $(\mathcal{L}, ^-, \mathcal{R}_s)$. Given a context $\mathcal{K}$,*

$\mathcal{N}_j = \quad (\mathcal{L}, ^-, \mathcal{R}_J \cup \mathcal{R}^c, \mathcal{K})$ *with* $\mathcal{R}^c = \bigcup_{s \in \mathcal{S}_0} \mathcal{R}_s^c$ *is a Jiminy argumentation theory.*

This induces a Jiminy argumentation framework $AF(\mathcal{N}_j) = (Arg(\mathcal{N}_j), Def(\mathcal{N}_j))$.

For any extension $\mathcal{E}$ of the full argumentation framework $AF(\mathcal{N}_\mathcal{S})$, we denote its restriction to the Jiminy framework by $\mathcal{E}^J = \mathcal{E} \cap Arg(\mathcal{N}_j)$ and call this set the *Jiminy fragment* of $\mathcal{E}$.

Definition 26 (Priority extension)**.** *Let AF be an argumentation framework and σ a semantics. We say that $\mathcal{E}$ is a priority extension of AF under σ if and only if (1) its Jiminy fragment $\mathcal{E}^J$ is an extension of the Jiminy framework $\mathcal{E}^J \in \sigma(AF(\mathcal{N}_j))$, and (2) $\mathcal{E}$ is compatible with respect to the priority information contained in $\mathcal{E}$. The set of priority extensions of AF under σ is denoted $\sigma^*(AF)$.*

7. In contrast to Brewka (1994), we do not require in Def. 26 that $\mathcal{E}$ is an extension of both the reduced framework $AF^\mathcal{E}$ *and* the original framework AF. The reason is that whenever the defeats Def and $Def^\mathcal{E}$ are incomparable in terms of $\subseteq$, so will be the extensions of AF and the extensions of $AF^\mathcal{E}$. In particular, the set of extensions of AF will be disjoint from those of any reduced framework $AF^{(\cdot)}$. See Example 4.20 below for an illustration of this.

Definition 27 (Jiminy dilemma checking and resolving)**.** *Let $DP = (mss, DV)$ be a decision problem, $\mathcal{N}_\mathcal{S} = (\mathcal{L}, \bar{}, \mathcal{R}_\mathcal{S}, \mathcal{K})$ an argumentation theory of $\mathcal{S}$ and σ a semantics. A fourth level (or $\mathcal{N}_\mathcal{S}$) dilemma for DP under σ is any moral dilemma in $\mathcal{C}_4 = \sigma^*(AF(\mathcal{N}_\mathcal{S}))$. That is, a dilemma exists if there are $\mathcal{E}_1 \in \sigma(AF^{\mathcal{E}_1}(\mathcal{N}_\mathcal{S}))$ and $\mathcal{E}_2 \in \sigma(AF^{\mathcal{E}_2}(\mathcal{N}_\mathcal{S}))$ such that*

$$(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV \text{ is inconsistent with respect to } \bar{}.$$

Otherwise, if for all $\mathcal{E}_1, \mathcal{E}_2 \in \sigma^*(AF(\mathcal{N}_{\mathcal{S}_0}))$, $(Obl(\mathcal{E}_1) \cup Obl(\mathcal{E}_2)) \cap DV = \emptyset$, then there is no dilemma at the fourth level and all third level dilemmas are resolved at the fourth level.

Example 4.18. Continue Example 4.17. $\mathcal{E}_1^+$ is a priority extension, i.e. $\mathcal{E}_1^+ \in pr^*(AF(\mathcal{N}_\mathcal{S}))$ since:

(1) its Jiminy fragment $\mathcal{E}_1^+ \cap Arg(\mathcal{N}_j) = \{W_1, \ldots, W_4, A_4, A_7, A_8\}$ is an extension of $AF(\mathcal{N}_j)$, and

(2) the priority $\succeq^{\mathcal{E}^+1}$ is compatible with $\mathcal{E}_1^+$.

In Example 4.17 we saw that condition (2) fails for the other candidates $\mathcal{E}_2^+, \mathcal{E}_3^+$. Thus, we only have one priority extension: $pr^*(AF(\mathcal{N}_\mathcal{S})) = \{\mathcal{E}_1^+\}$. From this, we conclude that all third level dilemmas are resolved at the fourth level. The set of obligations is $Obl(\mathcal{E}_1^+) = \{law, report, gdpr\}$, so the smart speaker decides to: { *comply with the law*, *report the potential threat*, *comply with the GDPR* }. (It refuses to *protect the user's privacy* but also *to collect data without explicit permission.*)

The following result suggests a simpler method for computing priority extensions.

Proposition 4.19. *Let $\mathcal{N}_\mathcal{S} = (\mathcal{L}, \bar{}, \mathcal{R}, \mathcal{K})$ be an argumentation theory of a set of stakeholders $\mathcal{S}$ and let $\sigma \in \{co, gr, pr, st\}$. For any priority extension $\mathcal{E} \in \sigma^*(AF(\mathcal{N}_\mathcal{S}))$, its Jiminy fragment $\mathcal{E}^J$ is a priority extension of the Jiminy framework: $\mathcal{E}^J \in \sigma^*(AF(\mathcal{N}_j))$.*

A simpler method for obtaining priority extensions is then: (1) find the priority extensions $\mathcal{E} = \mathcal{E}^J$ of the Jiminy framework $AF(\mathcal{N}_j)$; and (2) extend them with obligation and permission arguments into extensions $\mathcal{F}$ of the reduced framework: $\mathcal{F} \in \sigma(AF^\mathcal{E}(\mathcal{N}_\mathcal{S}))$. Since $\mathcal{E} = \mathcal{F}^J$, by Prop. 4.19, $\mathcal{F}$ will automatically be a priority extension: $\mathcal{F} \in \sigma^*(AF(\mathcal{N}_\mathcal{S}))$. Depending on the semantics σ, the expansion from $\mathcal{E}$ to some $\mathcal{F}$ will take one form or other:

($\sigma = gr$) Apply the algorithm for the grounded extension upon the priority extension $\mathcal{E}$.

($\sigma = co$) Fix a set of arguments that build from $\mathcal{E}$-arguments and are jointly conflict-free with this set. Close under defended arguments.

($\sigma = pr$) Proceed as in the complete semantics, but fix a maximal set of arguments that are built from and conflict-free with $\mathcal{E}$.

($\sigma = st$) Starting from $\mathcal{E}$, add one argument that is conflict-free with the set until the non-selected arguments are all defeated by this set.

Our definition of priority extension (Def. 26) differs from Brewka's original reduction (footnote 7). Let us motivate our weakened version by modifying once more the running example.

Example 4.20 (Running example without GDPR)**.** Remove from Ex. 4.2 the norm $i_1 \Rightarrow_L^r a_1$ for complying the GDPR. Hence, the argument A_6 that used it as its top norm no longer exists. This results in two preferred extensions $\mathcal{E}_2^+$ and $\mathcal{E}_3^+ \setminus \{A_6\}$, relative to Examples 4.16–4.17. These extensions differ from the unique extension of the reduced framework $\mathcal{E}_1^+ \setminus \{A_6\} \in pr(AF^{\mathcal{E}_1^+ \setminus \{A_6\}})$.

We thus obtain a priority extension $\mathcal{E}_1^+ \setminus \{A_6\} \in pr^*(AF)$, while under Brewka's original definition (fn. 7) no priority extension would exist.

Applying the four levels of dilemma checking and resolving. Depending on the application domain, the Jiminy system can make use of all four levels of argumentation or it can terminate at the earliest level without moral dilemmas and return the moral recommendation from this level.

for time-critical applications, such as self-driving vehicles, decisions must be made as quickly as possible. The argumentation system at level $i+1$ adds arguments (or defeats) to those of level i, resulting in an exponential increase in the number of candidate extensions (sets of arguments). For these applications, the Jiminy system can be implemented as an anytime algorithm: starting with level 1, it will keep returning better moral recommendations after reaching higher levels, as long as a prefixed deadline has not been met.

for sensitive applications, the stakeholders might agree upon any moral recommendation that is achieved at the earliest level that resolves all dilemmas; this might prevent applying the Jiminy norms to favour some stakeholders over the others unless strictly necessary (level 4), or enforce that one's norms are not used in combination with certain judgements (norms) from rival stakeholders (level 3).

In general, though, the highest the level of the argumentation framework, the better moral recommendations can be expected. Even in case that the moral recommendation remains the same between different levels, higher levels will provide with more comprehensive explanations for such output, i.e. based on more refined extensions and defeats.

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 will the following proposition.

Proposition 4.21. *Given a set of argumentation theories $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s, \mathcal{K})$ where $s \in \mathcal{S} = \{s_1, \ldots, s_n, J\}$, and a decision problem $DP = (mss, DV)$, the Jiminy will have one of the following two possible answers: there is a dilemma at level i, or there is no dilemma at level i and all dilemmas are resolved at level i, where $i = 1, 2, 3, 4$.*

Proof. According to Definitions 19, 20, 22 and 27, this proposition directly holds. □

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 & 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 & 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 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 one or several argumentation frameworks but also on the assessed level of the decision, and on whether Jiminy plays the role of ranking the stakeholders.

Consider Figures 5(a),(c) and Figure 6(d) again. In the reduced framework $AF(\mathcal{N}_{\mathcal{S}})$, 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" ($w_1 \Rightarrow^r_L d_1$). "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" ($w_3 \Rightarrow^r_H d_3$).

Explaining justification and rejection as a dialogue by referring to an argumentation graph. Argument A_1 is accepted because it has no defeater since the defeats $A_2 \rightarrow A_1$ and $A_5 \rightarrow A_1$ are removed by applying the priority relation encoded by the norms $w_3 \Rightarrow L \succ H$ and resp. $w_2 \Rightarrow L \succ M$ from the normative system of Jiminy. (Compare the defeats between Figures 6(a)–(c) and Figure 6(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 & 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 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 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 & Etzioni, 2017).

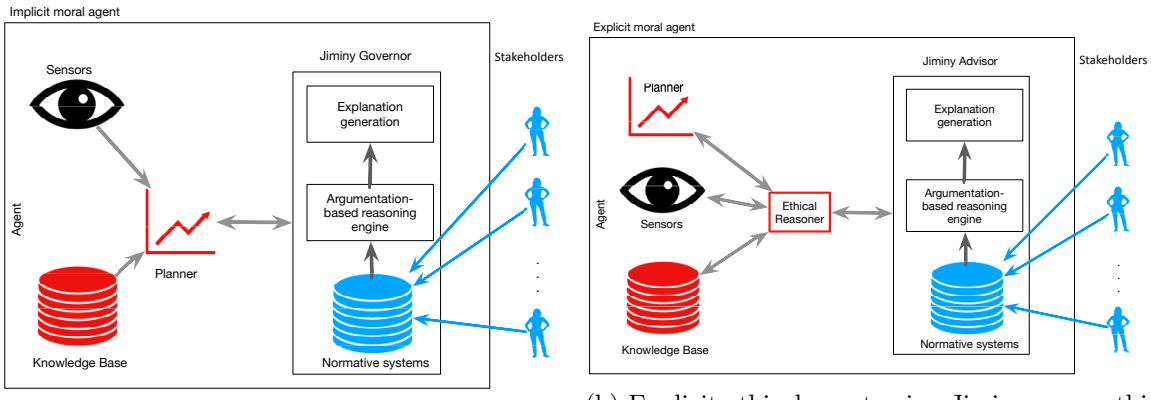
An *ethical-impact agent* does not make any ethically sensitive decisions itself and does not necessarily operate in ethically sensitive situations. However, by virtue of replacing some human activities with the artificial agent, we change the "moral environment" in which the agent operates. For example, 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 decisions or 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.



(a) Implicit ethical agent obtained by using Jiminy as an ethical governor

(b) Explicit ethical agent using Jiminy as an ethical advisor

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 7b). 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. In this paper we focused on specifying the subcomponents of its normative system, its argumentation reasoning engine, and its explanation generation engine. A running example was used to illustrate different aspects of these subcomponents.

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., 2018); 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 & Prakken, 2013) and some other work that extend ASPIC+, particularly by Baroni et al. (2015, 2018), 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, 2018) 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 depend on the types of norms involved, i.e., two permissive norms are never in conflict and institutional 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 & 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 & 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 "leasers" of the autonomous system[8]. 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), 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.

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 & 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

8. https://robohub.org/should-a-carebot-bring-an-alcoholic-a-drink-poll-says-it-depends-on-who-owns-the-robot/

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 or added to 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, 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.

Appendix: Proofs

This appendix contains the proofs most results found in Section 4.

Lemma 4.5 (Rationality postulates)**.** *The rationality postulates (Caminada & Amgoud, 2007) hold for the bcr-fragment $\mathcal{E}^{bcr}$ of any σ-extension $\mathcal{E}$ under $\sigma \in \{co, pr, gr, st\}$. In particular, direct consistency is satisfied by any bcr-fragment $\mathcal{E}^{bcr}$ of some $\mathcal{E} \in \sigma(AF)$; subargument closure holds for any extension $\mathcal{E} \in \sigma(AF)$ as well.*

Proof. (Direct consistency.) Recall any complete semantics σ is conflict free. Towards a contradiction, assume that some *bcr*-fragment $\mathcal{E}^{bcr} = \mathcal{E} \setminus \mathcal{A}^p$ of some $\mathcal{E} \in \sigma(AF)$ contains a pair $A, B \in \mathcal{E}$ such that $\text{conc}(A) \in \overline{\text{conc}(B)}$. By this assumption, A attacks B at B. (Case $A \succeq B$.) Then, we immediately have a direct defeat $(A, B) \in \mathit{Def}$, in contradiction with $\mathcal{E}$ being conflict free. (Case $A \prec B$.) Then A contains a subargument, namely A itself, attacking B at B and satisfying $B \succ A$, so a reverse defeat $(A, B) \in \mathit{Def}$ exists, again a contradiction. (Other cases.) From $A \not\succeq B$ and $A \not\prec B$, we have that A and B are incomparable, but by Def. 11 this is impossible, given that $A, B \in \mathcal{E}^{bcr} = \mathcal{E} \cap (\mathcal{A}^b \cup \mathcal{A}^c \cup \mathcal{A}^r)$.

(Subargument closure.) We show first that this postulate holds for extensions. Let $\mathcal{E} \in \sigma(AF)$ be an extension, and towards a contradiction let A, A' be arguments with $A' \in \text{sub}(A)$ and $A \in \mathcal{E}$ but $A' \notin \mathcal{E}$. Since σ is a complete semantics, any argument defended by $\mathcal{E} \in \sigma(AF)$ is in $\mathcal{E}$. Hence, from this and $A' \notin \mathcal{E}$, we infer that A' is not defended: $(\star)$ there is $B \in \mathit{Arg}$ defeating A' and such that $(C, B) \notin \mathit{Def}$ for any $C \in \mathcal{E}$. (Case 1.) Suppose first B attacks A' at some $A'' \in \text{Sub}(A')$ with $B \succeq A''$. Hence also B attacks A at A'' with $B \succeq A''$ and so B defeats A. Since $A \in \mathcal{E}$ and $\mathcal{E}$ does not defeat B, by $(\star)$ the defeat $(B, A) \in \mathit{Def}$ contradicts that $\mathcal{E}$ is admissible. (Case 2.) Suppose now that A' contains some $A'' \in \text{Sub}(A')$ attacking B at B and $B \succ A''$. From this and $A' \in \text{Sub}(A)$, it also holds that A contains $A'' \in \text{Sub}(A)$ that attacks B at B and satisfying $B \succ A''$. Again B defeats $A \in \mathcal{E}$ but by $(\star)$ $\mathcal{E}$ does not defeat B, in contradiction with $\mathcal{E}$ being admissible. This concludes the proof for extensions. Now, for the *bcr*-fragment $\mathcal{E}^{bcr} \subseteq \mathcal{E}$, since both sets $\mathcal{E}$ and $(\mathcal{A}^b \cup \mathcal{A}^c \cup \mathcal{A}^r)$ are closed under subarguments, so is their intersection $\mathcal{E}^{bcr}$.

The remaining rationality postulates (indirect consistency, closure under strict rules) trivially hold for this argumentation system, as our languages feature no strict rules. □

Proposition 4.7. *Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory with $\mathcal{R} = \mathcal{R}^c \cup \mathcal{R}^r \cup \mathcal{R}^p$. (1) For any extension $\mathcal{E} \in \sigma(AF(\mathcal{N}))$ under $\sigma \in \{co, pr, gr, st\}$ there exists (M^c, M^r, M^p), a norm extension of $(\mathcal{R}^c, \mathcal{R}^r, \mathcal{R}^p)$ in context $\mathcal{K}$, such that:*

(i) $I(M^c, \mathcal{K}) \supseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$;

(ii) $O(M^c, M^r, \mathcal{K}) \supseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^r)$;

(iii) $P(M^c, M^p, \mathcal{K}) \supseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^p)$.

(2) For the stable case $\sigma = st$, the inclusions in (i)–(iii) are in fact identities: (i') $I(M^c, \mathcal{K}) = \text{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$; (ii') $O(M^c, M^r, \mathcal{K}) = \text{Conc}(\mathcal{E} \cap \mathcal{A}^r)$; (iii') $P(M^c, M^p, \mathcal{K}) = \text{Conc}(\mathcal{E} \cap \mathcal{A}^p)$.

Proof. (1) Let $\sigma \in \{co, pr, gr, st\}$. We prove (i)–(iii) using that $\mathcal{E} \in \sigma(AF)$ implies that $\mathcal{E}$ defends itself and contains the arguments it defends. For $\tau \in \{c, r, p\}$, define:

$$M^\tau_\mathcal{E} = \big\{\phi \Rightarrow^\tau \psi \in \mathcal{R}^\tau :\ \exists A \in \mathcal{E}(A = A' \Rightarrow^\tau \psi \text{ and } \mathrm{conc}(A') = \phi)\big\}.$$

Let $\mathcal{E}^{bcr} \subseteq \mathcal{E}$ be the corresponding *bcr*-fragment. By direct consistency (Lemma 4.5), $\mathrm{Conc}(\mathcal{E}^{bcr})$ is consistent w.r.t. $^-$ and, as a consequence, so is each set $\mathrm{Conc}(\mathcal{E} \cap \mathcal{A}^\tau)$ with $\tau \in \{b, c, r\}$. We expand each set $M^\tau_\mathcal{E} \subseteq \mathcal{R}^\tau$ into a set M^τ, in the order *c-then-p-then-r*, and prove that (M^c, M^r, M^p) is a norm extension.

($\tau = c$.) Starting with $M'^c = M^c_\mathcal{E}$, we keep adding to the set M'^c, one rule at a time, a rule r from $\mathcal{R}^c \setminus M'^c$ that is triggered by $I(M'^c, \mathcal{K})$ and such that $I(M'^c \cup \{r\}, \mathcal{K}) \cup \mathrm{Conc}(\mathcal{E}^{bcr})$ is consistent. After this, we expand M'^c with all the remaining untriggered rules in $\mathcal{R}^c \setminus M'^c$. This defines M^c.

We check that M^c is a maximal subset of $\mathcal{R}^c$ such that $I(M^c, \mathcal{K})$ is consistent w.r.t. $^-$ (Def. 3). Suppose otherwise, so some $r = \psi \Rightarrow^c \phi$ in $\mathcal{R}^c$ exists such that r is triggered by $I(M^c, \mathcal{K})$ and $I(M^c \cup \{r\}, \mathcal{K})$ is consistent. By construction of M^c, there must be some $A \in \mathcal{E}^{bcr}$ such that $I(M^c \cup \{r\}, \mathcal{K}) \cup \{\mathrm{Conc}(A)\}$ is not consistent. By our assumption, this can only occur for some $A \in \mathcal{E} \cap \mathcal{A}^r$. Let then $B = B' \Rightarrow^c \phi$ be the argument built using r over some $B' \in \mathcal{E} \cap (\mathcal{A}^b \cup \mathcal{A}^c)$. By Def. 11, we have $A \prec B$ and so B defeats A. By admissibility, some $C \in \mathcal{E}$ defends A, with C attacking B at B, i.e. $\mathrm{Conc}(C) \in \overline{\phi}$; again, by Def. 11 we must have $C \in \mathcal{E} \cap (\mathcal{A}^b \cup \mathcal{A}^c)$. But since $\mathrm{Conc}(C) \in I(M^c_\mathcal{E}, \mathcal{K}) \subseteq I(M^c, \mathcal{K})$, we contradict the assumption that $I(M^c \cup \{r\}, \mathcal{K})$ was consistent.

It only remains to show the inclusion in (i). Clearly, $\mathcal{K} = I_0(M^c, \mathcal{K}) \subseteq I(M^c, \mathcal{K})$. Moreover, $I(M^c_\mathcal{E}, \mathcal{K})$ is consistent w.r.t. $^-$ and each step in the construction of M^c preserves this consistency; hence, $I(M^c, \mathcal{K})$ is also consistent. Let $r = \phi \Rightarrow^c \psi$ in $\mathcal{R}^\tau \setminus M^c$ be arbitrary. Thus, r is triggered by some element of $I(M^c, \mathcal{K})$, as otherwise we would have $r \in M^c$. If the addition of r was consistent with both $\mathrm{Conc}(\mathcal{E}^{bcr})$ and the consequents of M^c, then it would have been added to M^c, and so ψ and some such element would be contraries. (Base case) ψ is a contrary of some formula $\mathrm{Conc}(A)$ with $A \in \mathcal{E}^{bcr}$. Then, since r is triggered by M^c, an argument B in $\mathrm{Arg}(\mathcal{N})$ exists with $\mathrm{Conc}(B) = \phi$, and so the argument $C = B \Rightarrow^c \psi$ is also in $\mathrm{Arg}(\mathcal{N})$. Since C defeats A, by admissibility some argument $D \in \mathcal{E} \cap (\mathcal{A}^b \cup \mathcal{A}^c)$ must defend $A \in \mathcal{E}$, with D defeating C. Such D moreover can only attack C at C (since $\mathrm{Conc}(\mathcal{E})$ is consistent with all of M^c). Finally, since the rules of D are in $M^c_\mathcal{E} \subseteq M^c$ and the brute facts of D are in $\mathcal{K}$, we conclude that $I(M^c_\mathcal{E} \cup \{r\}, \mathcal{K})$ is inconsistent w.r.t. $^-$, and so is $I(M^c \cup \{r\}, \mathcal{K})$. (Ind. case) Suppose ψ is a contrary of the consequent of some triggered rule $r' \in M^c$. Then immediately $I(M^c \cup \{r\}, \mathcal{K})$ is inconsistent, since r' is triggered. This contradicts the construction of M^c.

($\tau = p$.) We just set $M^p = \mathcal{R}^p$. Let us show (iii). Let $A = A' \Rightarrow^p \phi$ be in $\mathcal{E}$ and let $r = \psi \Rightarrow^p \phi$ be the top norm of A. By the construction of A and the proof of Lemma 4.5 (subargument closure holds for extensions), $A' \in \mathcal{E} \cap (\mathcal{A}^c \cup \mathcal{A}^b)$, so we have that r is triggered by $I(M^c_\mathcal{E}, \mathcal{K}) \subseteq I(M^c, \mathcal{K})$. This and $M^p = \mathcal{R}^p$ imply that $\phi \in P(M^c, M^p, \mathcal{K})$.

($\tau = r$.) Starting with $M'^r = M^r_\mathcal{E}$ we add one rule r at a time from $\mathcal{R}^r \setminus M'^r$ that is triggered by $I(M^c, \mathcal{K})$ whenever $I(M^c, \mathcal{K}) \cup O(M^c, M'^r \cup \{r\}, \mathcal{K})$ is consistent w.r.t. $^-$ and all sets $O(M^c, M'^r \cup \{r\}, \mathcal{K}) \cup \{\phi\}$ with $\phi \in P(M^c, M^p, \mathcal{K})$ are consistent w.r.t. $^-$. After this, M^r is again defined by expanding M'^r with all the remaining rules in $\mathcal{R}^r$ not triggered by $I(M^c, \mathcal{K})$. Clearly, this construction leads to a subset $M^r \subseteq \mathcal{R}$ that is maximal with the two consistency conditions from Def. 3. Let us check the inclusion in (ii). Let $A = A' \Rightarrow^r \phi$ be in

$\mathcal{E}$ and let $r = \psi \Rightarrow^r \phi$ be the top norm of A. As before, $A' \in \mathcal{E} \cap \mathcal{A}^c$, so r is triggered by $I(M^c_{\mathcal{E}}, \mathcal{K})$. This and the fact that $r \in M^r_{\mathcal{E}} \subseteq M^r$ imply that $\phi \in O(M^c, M^r, \mathcal{K})$.

(2) For $\sigma = st$, we now prove that the inverse of the inclusions from (i)–(iii) hold for the stable semantics. This suffices for proving (i')–(iii') respectively. Let $\mathcal{E} \in st(AF(\mathcal{N}))$ and let $M = (M^c, M^r, \mathcal{R}^p)$ be defined as above.

(i') Let us show that $I(M^c, \mathcal{K}) \subseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$. The proof is by induction on the construction of M^c. (Base case) It is immediate that $I(M^c_{\mathcal{E}}, \mathcal{K}) \subseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$ from the definition of $M^c_{\mathcal{E}}$. (Ind. case) Suppose that for the construction so far of M^c, say as a set M'^c, it holds that $I(M'^c, \mathcal{K}) \subseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$. Let $r = \phi \Rightarrow^c \psi$ be the next (triggered) rule to be added to M'^c. We know that $I(M'^c \cup \{r\}, \mathcal{K})$ is consistent. Since r is triggered, let $A \in (\mathcal{A}^b \cup \mathcal{A}^c)$ be such that $\text{Conc}(A) = \phi$, and let $B = A \Rightarrow^c \psi$. If $B \in \mathcal{E}$, we are done. Otherwise, $B \notin \mathcal{E}$ implies (by stability) that $\mathcal{E}$ defeats B, say with an argument $C \in \mathcal{E}$ that (by Def. 11) is also in $\mathcal{A}^b \cup \mathcal{A}^c$. But this is impossible, since by construction of M^c, all the triggered rules in M^c are consistent with $\text{Conc}(\mathcal{E}^{bcr})$, and so such an attack cannot exist.

(ii') We prove now that $O(M^c, M^r, \mathcal{K}) \subseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^r)$. Let $r = \psi \Rightarrow^c \phi$ be a rule in M^r triggered by $I(M^c, \mathcal{K})$. Using this and (i'), we have that $\psi \in I(M^c, \mathcal{K}) \subseteq \text{Conc}(\mathcal{E} \cap \mathcal{A}^c) \cup \mathcal{K}$, so let $A' \in \mathcal{E} \cap \mathcal{A}^c$ be an argument for such ψ. Define $A = A' \Rightarrow^r \phi$. Clearly, $A \in \mathcal{A}^r$, and if $A \in \mathcal{E}$ we are done, so assume otherwise towards a contradiction. Again $A \notin \mathcal{E}$ implies there is $B \in \mathcal{E}$ such that $(B, A) \in \mathit{Def}(\mathcal{N})$. (Case $B \in \mathcal{A}^b \cup \mathcal{A}^c$.) Impossible, since M^r was defined (for triggered rules) as a set of rules consistent with $I(M^c, \mathcal{K})$ and the latter set contains only conclusions from $\mathcal{E}$. (Case $B \in \mathcal{A}^r$.) Again, the inductive construction of M^r makes this case impossible, since $B \in \mathcal{E}^{bcr}$ and the inconsistency of the triggered r with B would imply that $r \notin M^r$. (Case $B \in \mathcal{A}^p$.) Let $B = B' \Rightarrow^p \phi'$ be built over some rule $r' = \psi' \Rightarrow^p \phi'$ in $\mathcal{R}^p$ with $\phi' \in \overline{\phi}$ or $\phi \in \overline{\phi'}$. By subargument closure and Def. 11, $B' \in \mathcal{E} \cap (\mathcal{A}^b \cup \mathcal{A}^c)$. Moreover, by (i') r' is triggered by $I(M^c, \mathcal{K})$, and so $\phi' \in P(M^c, \mathcal{R}^p, \mathcal{K})$. But this contradicts the construction of M^r during the addition of r, since $O(M^c, M'^r \cup \{r\}, \mathcal{K}) \cup \{\phi'\}$ is not consistent for some $\phi' \in P(M^c, M^p, \mathcal{K})$.

(iii') Let $\phi \in P(M^c, \mathcal{R}^p, \mathcal{K})$. We show that $\phi \in \text{Conc}(\mathcal{E} \cap \mathcal{A}^p)$. As before, let $r = \psi \Rightarrow^p \phi$ be a triggered rule in $\mathcal{R}^p$, i.e. with $\psi \in I(M^c, \mathcal{K})$. By (i'), some $A' \in \mathcal{E} \cap (\mathcal{A}^b \cup \mathcal{A}^c)$ exists with $\text{Conc}(A') = \psi$. Let then $A = A' \Rightarrow^p \phi$. By Def. 11, there can be no argument defeating A at A, so from this and $A' \in \mathcal{E}$ we conclude that also $A \in \mathcal{E}$. Finally, $\phi \in \text{Conc}(\mathcal{E} \cap \mathcal{A}^p)$. □

Proposition 4.8. *Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory with a symmetric contrariness function $^-$. Then, any P-maximal norm extension $M = (M^c, M^r, \mathcal{R}^p)$ in $\mathcal{K}$ contains a σ-extension $\mathcal{E}$ in the sense of the (i)–(iii) from Prop. 4.7 for any $\sigma \in \{co, pr, gr\}$.*

Proof. Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory with a symmetric function $^-$ and let $AF = (\mathit{Arg}(\mathcal{N}), \mathit{Def}(\mathcal{N}))$ be induced by $\mathcal{N}$. Let also $M = (M^c, M^r, \mathcal{R}^p)$ be a P-maximal norm extension in $\mathcal{K}$ and $\sigma \in \{co, pr, gr\}$. Define $\mathcal{N}_M = (\mathcal{L}, ^-, M^c \cup M^r \cup \mathcal{R}^p, \mathcal{K})$. Then it suffices to check that $\mathit{Arg}(\mathcal{N}_M)$ forms a *co*-extension in AF, namely $\mathcal{E}_M = \mathit{Arg}(\mathcal{N}_M) \subseteq \mathit{Arg}(\mathcal{N})$. (From this, minimal and maximal *co*-extensions will exist as well for $\sigma = gr$ and resp. $\sigma = pr$.) To this end, we show that $\mathcal{E}_M \in co(AF)$.

(*Conflict free.*) Suppose towards a contradiction that $(A, B) \in \mathit{Def}(\mathcal{N})$ for some $A, B \in \mathit{Arg}(\mathcal{N}_M)$. Since the latter set is closed under subarguments, we can assume w.l.o.g. that A attacks B at B and either $A \succeq B$ (direct defeat) or $A \prec B$ (reverse defeat). Note

that all the rules occurring in A, B are triggered by $I(M^c, \mathcal{K})$. Based on this, it can be checked that for any possible pair $\tau, \tau' \in \{b, c, r, p\}$ satisfying $A \in \mathcal{A}^\tau$ and $B \in \mathcal{A}^{\tau'}$, one of the consistency conditions for the sets $I(M^c, \mathcal{K}) \cup O(M^c, M^r, \mathcal{K})$, or $O(M^c, M^r, \mathcal{K}) \cup \{\phi\}$ for some $\phi \in P(M^c, \mathcal{R}^p, \mathcal{K})$, is violated (by the top rules of A, B being in M). But this contradicts that M is a norm extension.

(*Admissibility.*) Let $(B, A) \in \mathit{Def}(\mathcal{N})$ be such that $B \in \mathit{Arg}(\mathcal{N}) \setminus \mathcal{E}_M$ and $A \in \mathcal{E}_M$. As before, and by the symmetry of $^-$, we can just assume that B attacks A at A and $B \succeq A$.

(Case $B \in \mathcal{A}^b$.) Impossible since then $A, B \in \mathcal{E}_M^{bcr}$ is in contradiction with the direct consistency of $\mathcal{E}_M^{bcr}$ from Lemma 4.5.

(Case $B \in \mathcal{A}^c$.) From $B \notin \mathcal{E}_M$ and the maximality of M^c, there is either a triggered rule $r = \psi \Rightarrow^c \phi \in M^c$ or simply $\{\phi\} \in \mathcal{A}^b$ such that $\phi \in \overline{\mathrm{Conc}(B)}$ (by the symmetry of $^-$ we can ignore the opposite $\mathrm{Conc}(B) \in \overline{\phi}$). In either case, there is an argument $C = C' \Rightarrow^c \phi$ or resp. $C = \{\phi\}$ in $\mathcal{E}_M$. In the former case, we have $C' \Rightarrow^c \phi \in \mathcal{A}^c$ and so $C \succeq B$; in the latter, $C \succ B$. In any case, this argument C satisfies $(C, B) \in \mathit{Def}(\mathcal{N})$.

(Case $B \in \mathcal{A}^r$.) By Def. 11, we have that also $A \in \mathcal{A}^r$. Let $B = B' \Rightarrow^r \phi$ for some $r = \psi \Rightarrow^r \phi$. By symmetry, also A attacks B at B and $A \succeq B$. Hence, $(A, B) \in \mathit{Def}(\mathcal{N})$.

(Case $B \in \mathcal{A}^p$.) By Def. 11, $A \in \mathcal{A}^r$. And, by symmetry, let $B = B' \Rightarrow^p \phi$ be built using some rule $r = \psi \Rightarrow^p \phi$ with $\phi \in \overline{\mathrm{Conc}(A)}$. We show first that r cannot be triggered by $I(M^c, \mathcal{K})$. Suppose the contrary: then we have $B' \in \mathcal{E}_M$ and, with $r \in M^p = \mathcal{R}^p$, conclude that $\phi \in P(M^c, M^p, \mathcal{K})$. But this contradicts that $O(M^c, M^r, \mathcal{K}) \cup \{\phi\}$ is consistent for this particular ϕ. Now, from r not being triggered we obtain that $B' \notin \mathcal{E}_M$. Hence, also some $\mathcal{R}^c$-rule occurring in B' is not triggered by $I(M^c, \mathcal{K})$. Let $r' = \theta \Rightarrow^c \psi'$ be the earliest such rule in B'; that is, let r' occur as a top rule of a subargument $C = C' \Rightarrow^c \psi'$ of B' such that $C' \in \mathcal{E}_M$. Then, by the maximality of M^c with the consistency of $I(M^c, \mathcal{K})$, this set $I(M^c, \mathcal{K})$ contains (by symmetry) a contrary, say $\alpha \in \overline{\psi'}$. In other words, M^c contains a triggered rule r'' of the form $\ldots \Rightarrow^c \alpha$, and so an argument $D = \ldots \Rightarrow^c \alpha$ exists in $\mathcal{E}_M$ that attacks C at C. Since $C', D \in \mathcal{A}^c$, we have $D \succeq C$ and thus conclude that $(D, C) \in \mathit{Def}(\mathcal{N})$.

In summary, no matter the nature of B, $\mathcal{E}_M$ defends $A \in \mathcal{E}_M$.

(*Closure under defended arguments.*) The claim is trivial for brute fact (defended) arguments $A = \{\phi\}$, so suppose $\mathcal{E}_M$ defends an argument $A \in \mathit{Arg}(\mathcal{N})$ of the form $A = A' \Rightarrow^\tau \phi$. Towards a contradiction, suppose that $A \notin \mathcal{E}_M$. Let, moreover, such A be minimal with this property: if A' (or any other subargument) is defended by $\mathcal{E}_M$, then $A' \in \mathcal{E}_M$. Since, indeed, A' is defended by $\mathcal{E}_M$ (provided that A is), we obtain that $A' \in \mathcal{E}_M$. Since $A \notin \mathcal{E}_M$ but $A' \in \mathcal{E}_M$, the top rule of A, say $r = \ldots \Rightarrow^\tau \phi$, is not in M^τ. And since $A' \in \mathcal{E}_M$, the rule r is triggered by $I(M^c, \mathcal{K})$, and so by the maximality of M^τ there must be a rule $r' = \ldots \Rightarrow^{\tau'} \theta$ in $M^{\tau'}$ with $\theta \in \overline{\phi}$ and that is also triggered by $I(M^c, \mathcal{K})$. The latter implies that an argument $B = B' \Rightarrow^{\tau'} \theta$ with top rule r' exists in $\mathit{Arg}(\mathcal{N})$ and is such that $B' \in \mathcal{E}_M$. Note that B is a defeater of A at A.

(Case $\tau' = r$.) Then, a defeater C of B at B, is either a brute fact argument (in contradiction with r' being in $M^{\tau'}$), or it is built with top rule $r'' \in M^c \cup M^r$, say $r'' = \ldots \Rightarrow^c \psi$ or $r'' = \ldots \Rightarrow^r \psi$ for some $\psi \in \overline{\theta}$. In either case, the fact $r'' \in M$ contradicts that $I(M^c, \mathcal{K}) \cup O(M^c, M^r, \mathcal{K})$ is consistent since $r' \in M^r$.

(Case $\tau' = c$.) The proof is similar, with only the two cases: $C \in \mathcal{A}^b$, or r'' a rule in M^c.

(Case $\tau' = p$.) A defeater C of B at B cannot exist, contradicting that $\mathcal{E}_M$ defends A.

In all cases of τ' we reached a contradiction, so $r \in M^\tau$ and, finally, $A \in \mathcal{E}_M$. □

We compare again a P-maximal norm extension $M = (M^c, M^r, \mathcal{R}^p)$ and the argumentation theory induced by it: $\mathcal{N}_M = (\mathcal{L}, ^-, M^c \cup M^r \cup \mathcal{R}^p, \mathcal{K})$.

Proposition 4.9. *Let $\mathcal{N} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory, inducing the argumentation framework $AF = (Arg(\mathcal{N}), Def(\mathcal{N}))$. (1) For any naive extension $\mathcal{E} \in na(AF)$, it holds that $\mathcal{E} = Arg(\mathcal{N}_M)$ for some norm extension M in $\mathcal{K}$. (2) For any P-maximal norm extension M in $\mathcal{K}$, the set $\mathcal{E}_M = Arg(\mathcal{N}_M)$ is a naive extension: $\mathcal{E}_M \in na(AF)$.*

Proof. (1) Let $\mathcal{E} \in na(AF)$. We define $M = (M^c, M^r, M^p)$. For each $\tau \in \{c, r, p\}$, let

$$M^\tau = \bigcup_{A \in \mathcal{E}} (\text{Rules}(A) \cap \mathcal{R}^\tau) \cup \{\psi \Rightarrow^\tau \phi \in \mathcal{R}^\tau : \psi \notin \text{Conc}(\mathcal{E} \cap \mathcal{A}^c)\}.$$

Let us check that M is a (P-maximal) norm extension in $\mathcal{K}$. (Consistency.) Clearly, each of the sets $I(M^c, \mathcal{K})$ and $I(M^c, \mathcal{K}) \cup O(M^c, M^r, \mathcal{K})$ and $O(M^c, M^r, \mathcal{K}) \cup \{\phi'\}$ for some $\phi' \in P(M^c, M^p, \mathcal{K})$ is consistent, as otherwise a defeat would occur within $\mathcal{E}$, namely inside $\mathcal{E} \cap \mathcal{A}^c$ or $\mathcal{E} \cap (\mathcal{A}^c \cup \mathcal{A}^r)$ or resp. $\mathcal{E} \cap (\mathcal{A}^r \cup \mathcal{A}^p)$. (Maximal consistency.) Suppose towards a contradiction that some of the sets M^τ is not maximal with this property, i.e. suppose that all the above sets $I(\cdot,\cdot), \ldots, O(\cdot,\cdot) \cup \{\phi'\}$ are also consistent when we add some rule $r \in \mathcal{R}^\tau$ to M^τ, say $r = \psi \Rightarrow^\tau \phi$.

(Case $\tau = c$.) From $r \notin M^c$ and the definition of M^c, we obtain that r is triggered: $\psi \in \text{Conc}(\mathcal{E} \cap \mathcal{A}^c)$ and so $\psi \in I(M^c, \mathcal{K})$. Thus an argument $A' = \ldots \Rightarrow^c \psi$ or $A' = \{\psi\}$ exists in $\mathcal{E}$, from which we can build the argument $A = A' \Rightarrow^c \phi$. Since the addition of r to M^c preserves the consistency of the set $I(M^c \cup \{r\}, \mathcal{K}) \cup O(M^c, M^r, \mathcal{K})$, the argument A neither attacks nor is attacked by any element of $\mathcal{E} \cap (\mathcal{A}^c \cup \mathcal{A}^r)$; in addition, neither a defeat from/to $\mathcal{E} \cap \mathcal{A}^p$ can exist. Thus, $\mathcal{E} \cup \{A\}$ is conflict free, so from the assumption $\mathcal{E} \in na(AF)$, we conclude that $A \in \mathcal{E}$, which now implies that $r \in M^\tau$ and thus contradicts that $r \notin M^\tau$.

(Case $\tau = r$.) The reasoning is analogous, now adding some rule $r \in \mathcal{R}^r \setminus M^r$ and using the consistency of $I(M^c, \mathcal{K}) \cup O(M^c, M^r \cup \{r\}, \mathcal{K})$ and all sets of the form $O(M^c, M^r \cup \{r\}, \mathcal{K}) \cup \{\phi'\}$ to exclude any defeat involving the new argument $A = A' \Rightarrow^r \phi$ and $\mathcal{E}$.

(Case $\tau = p$.) Since permission arguments only attack obligation arguments, the proof is as in the previous case. Because of this, the above definition of M^p implies that $M^p = \mathcal{R}^p$.

(2) We check that $Arg(\mathcal{N}_M)$ is a naive extension whenever $M = (M^c, M^r, \mathcal{R}^p)$ is a norm extension. Let $\mathcal{E} = Arg(\mathcal{N}_M)$. We prove by induction on the complexity of an arbitrary $A \in Arg(\mathcal{N})$ that if $\mathcal{E} \cup \{A\}$ is conflict free, then $A \in \mathcal{E}$. (Base case.) For brute fact arguments A, clearly all of them are conflict free and in $\mathcal{E}$. (Inductive case.) Assume that all subarguments A' of some $A \in Arg(\mathcal{N})$ satisfy: if $\mathcal{E} \cup \{A'\}$ is conflict free, then $A' \in \mathcal{E}$. Now suppose that also $\mathcal{E} \cup \{A\}$ is conflict free. We show that $A \in \mathcal{E}$. Let $A = A' \Rightarrow^\tau \phi$ be built using some top rule $r = \psi \Rightarrow^\tau \phi'$. Since $\mathcal{E} \cup \{A\}$ is conflict free, so is $\mathcal{E} \cup \{A'\}$ and so $A' \in \mathcal{E}$. Assume, towards a contradiction, that $A \notin \mathcal{E}$. Hence, since r is triggered by M, $r \notin M^\tau$. (Hence $\tau \neq p$, since $M^p = \mathcal{R}^p$.) Thus, by the maximality of M^τ with the corresponding consistency condition, one of the following sets is inconsistent if we add r to M^τ: $I(M^c \cup \{r\}, \mathcal{K}) \cup O(M^c \cup \{r\}, M^r, \mathcal{K})$ (if $\tau = c$); or, if $\tau = r$, either $I(M^c, \mathcal{K}) \cup O(M^c, M^r \cup \{r\}, \mathcal{K})$ or $O(M^c, M^r \cup \{r\}, \mathcal{K}) \cup \{\phi'\}$ for some $\phi' \in P(M^c, \mathcal{R}^p, \mathcal{K})$. In any of these cases it can be checked that a defeat exists between A and some $B \in \mathcal{A}^c$ or resp. between A and some $B \in \mathcal{A}^r \cup \mathcal{A}^p$. Since for any such B, $B \in \mathcal{E}$, we reached a contradiction with the assumption that $\mathcal{E} \cup \{A\}$ was conflict free. □

Proposition 4.19. *Let $\mathcal{N}_{\mathcal{S}} = (\mathcal{L}, ^-, \mathcal{R}, \mathcal{K})$ be an argumentation theory of a set of stakeholders $\mathcal{S}$ and let $\sigma \in \{co, gr, pr, st\}$. For any priority extension $\mathcal{E} \in \sigma^*(AF(\mathcal{N}_{\mathcal{S}}))$, its Jiminy fragment $\mathcal{E}^J$ is a priority extension of the Jiminy framework: $\mathcal{E}^J \in \sigma^*(AF(\mathcal{N}_j))$.*

Proof. Let us abbreviate $AF(\mathcal{N}_{\mathcal{S}}) = (Arg(\mathcal{N}_{\mathcal{S}}), Def(\mathcal{N}_{\mathcal{S}}))$ as $AF = (Arg, Def)$. By Def. 11,

($\star$) all defeats of arguments in $Arg(\mathcal{N}_j)$ are from arguments in this set $Arg(\mathcal{N}_j)$; in other words, $Def' \cap (Arg(\mathcal{N}_{\mathcal{S}}) \times Arg(\mathcal{N}_j)) = Def' \cap (Arg(\mathcal{N}_j) \times Arg(\mathcal{N}_j))$ for any $Def' \in \{Def, Def^{\mathcal{E}}, \ldots\}$.

(Case $\sigma = co$.) Let $\mathcal{E} \in co^*(AF(\mathcal{N}_{\mathcal{S}}))$ and $A \in Arg$ be arbitrary. That is, (1) $\mathcal{E}^J \in co(AF(\mathcal{N}_j))$ and (2) $\mathcal{E} \in co(AF^{\mathcal{E}})$. We have to show that:

$$\text{(1')}\ \mathcal{E}^J \in co(AF(\mathcal{N}_j)) \qquad \text{and} \qquad \text{(2')}\ \mathcal{E}^J \in co(AF^{\mathcal{E}^J}),$$

But (1') is just (1), while (2') is equivalent to: $\mathcal{E}^J \in co(AF^{\mathcal{E}})$, so we prove the latter. Using (2), $A \in \mathcal{E} \Leftrightarrow \mathcal{E}$ defends A (in $AF^{\mathcal{E}}$), i.e. for all $(B, A) \in Def^{\mathcal{E}}$, there is $(C, B) \in Def^{\mathcal{E}}$. Now consider $AF^{\mathcal{E}}(\mathcal{N}_j) = (Arg(\mathcal{N}_j), Def(\mathcal{N}_j))$. Clearly $\mathcal{E}^J$ is conflict-free since $\mathcal{E}$ is. It remains to show the above $\Leftrightarrow$-equivalence for $\mathcal{E}^J$ and an arbitrary $A \in Arg(\mathcal{N}_j)$. ($\Leftarrow$) If $\mathcal{E}^J$ defends A (in $AF^{\mathcal{E}}(\mathcal{N}_j)$), then by $\mathcal{E}^J \subseteq \mathcal{E}$ and ($\star$) also $\mathcal{E}$ does defend A (in $AF^{\mathcal{E}}$). By (2), $A \in \mathcal{E}$ and finally by def. of $\mathcal{E}^J$, it also holds that $A \in \mathcal{E}^J$. ($\Rightarrow$) Suppose $A \in \mathcal{E}^J$. Hence $A \in \mathcal{E}$. By (2), we have that $\mathcal{E}$ defends A (in $AF^{\mathcal{E}}$). By ($\star$), we obtain that $\mathcal{E}^J$ also defends A.

(Case $\sigma = gr$.) Let $\mathcal{E} \in gr^*(AF)$. Thus, (1) $\mathcal{E}^J$ is $\subseteq$-minimal within $AF(\mathcal{N}_j)$ w.r.t. the properties (a) conflict-free, (b) defending itself and (c) the closure under defended arguments; and (2) $\mathcal{E}$ is $\subseteq$-minimal within $AF^{\mathcal{E}}$ w.r.t. (a)–(c). Since the Jiminy fragment of $\mathcal{E}^J$ is $\mathcal{E}^J$ itself it obviously satisifes (a)–(c) w.r.t. the same framework $AF(\mathcal{N}_j)$. It only remains to show that (2') $\mathcal{E}^J$ is in $gr(AF^{E^J}(\mathcal{N}_j)) = gr(AF^{\mathcal{E}}(\mathcal{N}_j))$. Clearly, $\mathcal{E}^J \in \lfloor\wr(AF(\mathcal{N}_j))$: (a) $\mathcal{E}^J$ is conflict-free since so is $\mathcal{E}$; (b) that $\mathcal{E}^J$ defends itself follows from the fact that $\mathcal{E}$ defends itself using ($\star$); (c) to see that $\mathcal{E}^J$ is closed under defended arguments, any defended argument in $Arg(\mathcal{N}_j)$ would also be defended by $\mathcal{E}$ given ($\star$), hence it would belong to $\mathcal{E}$ and then $\mathcal{E}^J$. Finally suppose, towards a contradiction, that $\mathcal{E}^J$ is not $\subseteq$-minimal with these properties, so there is some $\mathcal{E}^{J-} \subsetneq \mathcal{E}^J$ that satisifes (a)–(c) as well. Then define a set $\mathcal{E}^- = \mathcal{E} \setminus \bigcup_n \mathcal{F}_n$ by removing from $\mathcal{E}$ the following inductive construction: $\mathcal{F}_0 =$ the set of arguments in or built from $\mathcal{E}^J \setminus \mathcal{E}^{J-}$; and $\mathcal{F}_{n+1} =$ the set of arguments defended only by arguments in $\bigcup_{m \leq n} \mathcal{F}_m$. It can be checked that: $\mathcal{E}^-$ is a proper subset of $\mathcal{E}$ (since $\mathcal{E}^{J-} \subsetneq \mathcal{E}^J$); and also that $\mathcal{E}^-$ is a complete extension. This contradicts the initial assumption (2) that $\mathcal{E} \in gr(AF^{\mathcal{E}})$.

(Case $\sigma = pr$.) Let $\mathcal{E} \in pr^*(AF)$, so that (1) $\mathcal{E}^J \in pr(AF(\mathcal{N}_j))$ and (2) $\mathcal{E} \in pr(AF^{\mathcal{E}})$. By (1), it only remains to prove (2') $\mathcal{E}^J \in pr(AF^{\mathcal{E}^J}) = pr(AF^{\mathcal{E}})$. As in the previous proof for $\sigma = gr$, $\mathcal{E}^J \in co)(AF^{\mathcal{E}})$. So suppose towards a contradiction that $\mathcal{E}$ is not $\subseteq$-maximal with (a) conflict-freeness, (b) defending itself and (c) closure under defended arguments. Thus, there is some $A \in Arg(\mathcal{N}_j) \setminus \mathcal{E}^J$ such that $\mathcal{E}^J \cup \{A\}$ satisfies (a)–(c). Define then $\mathcal{F}_0 = \mathcal{E}^J \cup \{A\}$ and $\mathcal{F}_{n+1} =$ the set of arguments in Arg defended by $\bigcup_{m \leq n} \mathcal{F}_m$. Then the set $\mathcal{F} = \bigcup_n \mathcal{F}_m$ satisfies (a)–(c) and properly extends $\mathcal{E}$ with (at least) A, in contradiction with the assumption $\mathcal{E} \in pr(AF^{\mathcal{E}})$.

(Case $\sigma = st$.) Let $\mathcal{E} \in st^*(AF)$. That is, (1) $\mathcal{E}^J \in st(AF(\mathcal{N}_j))$ and (2) $\mathcal{E}$ is conflict-free and for any $A \in Arg$, it holds that $A \in Arg \setminus \mathcal{E} \Leftrightarrow (B, A) \in Def^{\mathcal{E}}$ for some $B \in \mathcal{E}$. We need to prove that (1') $\mathcal{E}^J \in st(AF(\mathcal{N}_j))$, which is just (1), and that (2') $\mathcal{E}^J \in st(AF^{\mathcal{E}^J}(\mathcal{N}_j))$ or

equivalently, that $\mathcal{E}^J$ is also conflict-free (which is immediate) and $\mathcal{E}^J \in st(AF^{\mathcal{E}^J}(\mathcal{N}_j))$. We prove a similar $\Leftrightarrow$ equivalence as the one above. ($\Rightarrow$). Suppose $A \in Arg(\mathcal{N}_j) \setminus \mathcal{E}^J$. By the definition of these two sets, we also have $A \in Arg(\mathcal{N}_j) \setminus \mathcal{E}$. By (2), $\mathcal{E}$ defeats A (in Arg); and finally by ($\star$) $\mathcal{E}^J$ also defeats A. ($\Leftarrow$). Now suppose that $\mathcal{E}^J$ also defeats some $A \in Arg(\mathcal{N}_j)$. Clearly, $\mathcal{E}$ defeats A as well, and by (2) $A \in Arg \setminus \mathcal{E}$. Finally, the latter fact and $A \in Arg(\mathcal{N}_j)$ and the def. of $\mathcal{E}^J$ together imply that $A \notin \mathcal{E}^J$. □

Acknowledgments

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/DLAl/van der Torre).

References

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

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

Anderson, M., & Leigh Anderson, S. (2014). GenEth: A general ethical dilemma analyzer. In *Proceedings of the 28th AAAI Conference on AI*, pp. 253–261.

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.

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.

Atkinson, K., & Bench-Capon, T. J. M. (2018). Taking account of the actions of others in value-based reasoning. *Artif. Intell.*, *254*, 1–20.

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.

Baroni, P., Gabbay, D., Giacomin, M., & van der Torre, L. (Eds.). (2018). *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.

Baroni, P., Giacomin, M., & Liao, B. (2018). A general semi-structured formalism for computational argumentation: Definition, properties, and examples of application. *Artif. Intell.*, *257*, 158–207.

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & SOCIETY*, *35*(1), 165–176.

Bench-Capon, T. J. M., & Modgil, S. (2017). Norms and value based reasoning: justifying compliance and violation. *Artif. Intell. Law*, *25*(1), 29–64.

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., Håvik, 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*, pp. 23–29. ACM.

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, pp. 303–319. Springer.

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*, pp. 453–454. IOS Press.

Brandt, F., Conitzer, V., Endriss, U., Lang, L., & Procaccia, A. D. (Eds.). (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

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.

Brewka, G., & Eiter, T. (1999). Preferred answer sets for extended logic programs. *Artif. Intell.*, *109*(1-2), 297–356.

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.

Caminada, M., & Amgoud, L. (2007). On the evaluation of argumentation formalisms. *Artif. Intell.*, *171*(5-6), 286–310.

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*.

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*, pp. 345–373. College Publications.

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.

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the 26th IIJCAI*, pp. 4698–4704.

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.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*(4), 403–418.

Fan, X., & Toni, F. (2015). On computing explanations in argumentation. In *AAAI*, pp. 1496–1502. AAAI Press.

Gabbay, D., Horty, 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.

Grossi, D., & Jones, A. (2013). Constitutive norms and counts-as conditionals. In Horty, J., Gabbay, D., Parent, X., van der Meyden, R., & van der Torre, L. (Eds.), *Handbook of Deontic Logic and Normative Systems*, pp. 407–441. College Publications, London.

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*.

Horty, J. F. (1994). Moral dilemmas and nonmonotonic logic. *J. Philos. Log.*, *23*(1), 35–65.

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

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.

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*, pp. 29–54. IOS Press.

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, pp. 117–124. ACM.

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.

Modgil, S., & Prakken, H. (2013). A general account of argumentation with preferences. *Artif. Intell.*, *195*, 361–397.

Moor, J. H. (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.

Reiter, R. (1980). A logic for default reasoning. *Artif. Intell.*, *13*(1-2), 81–132.

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*, p. 209. ACM.

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.

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.

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.

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

Wallach, 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.