Advanced Intelligent Systems and Reasoning: Standardization, Experimentation, Explanation

Pere Pardo[1] Leendert van der Torre[1,2] Liuwen Yu[1]

[1] *University of Luxembourg, Luxembourg*
[2] *Zhejiang University, China*

Abstract

We offer a perspective on advanced intelligent systems and reasoning, using as an example morally-decisive robots, as proposed in machine ethics. Given that norms often conflict, formal methods are necessary to resolve these conflicts in order to make morally acceptable or optimal decisions. The underlying basis of current algorithms spans from logical representation and reasoning to machine learning algorithms. We explore multiple methodologies including deontic ASP for standardizing normative reasoning, LogiKEy for testing ethical and legal reasoners, and formal argumentation for achieving explanatory transparency. Our vision is demonstrated using the argumentation-based Jiminy moral advisor. We also hint at future work that situates 'real-world' dialogue exchanges as the forum for discussing moral decisions, and we discuss the development of a platform for experimental user studies at the Zhejiang University – University of Luxembourg Joint Lab on Advanced Intelligent Systems and REasoning (ZLAIRE).

Keywords: Artificial intelligence, knowledge representation and reasoning, logic, formal argumentation, deontic answer set programming, LogiKEy, normative multiagent systems, machine ethics

1 Introduction

Our future, as much as it is a projection of the present, is also a reflection of the narratives we create, especially those crafted in the realm of science fiction. This genre, an intriguing amalgamation of philosophy and speculative thinking, serves as a canvas for portraying potential advancements in technology. A paramount example of such advancements is the development of advanced intelligent reasoners, artificial intelligence (AI) systems that encapsulate philosophical concepts such as rationalism and empiricism.

Standardization plays a fundamental role in the development of these AI systems. By ensuring consistency and predictability, it enables meaningful scientific experiments and allows us to gather empirical data. This process enriches our understanding of these advanced systems, and it expands our comprehension of reality, creating parallels with speculative narratives of science fiction.

Historically, the collective imagination has often cast AI in the mold of robotics. However, the reality in the coming years will deviate from this norm. While the world will not teem with robots as science fiction might suggest, we will witness the marked presence of AI. This visibility will not be in the form of physical machines but rather the rapid evolution and maturation of AI software. In a few years, core fields like computer vision, machine learning and human-machine interaction will have matured and will become integral to computer science technology.

The coming decade is set to mark a significant shift in the focus of AI. After conquering basic aspects of animal and infant intelligence, attention will turn towards adult-level human intelligence. This new focus will entail an understanding of knowledge representation, interaction with other agents, and grappling with ethical, legal, and social systems. These advances will bring to the fore two main challenges: individual reasoning and collective reasoning.

Individual reasoning involves theoretical and practical reasoning, whereas collective reasoning delves into multiagent dialogues and collaborations. Navigating the balance and interplay between these two types of reasoning will be of central concern.

The Zhejiang University – University of Luxembourg Joint Lab on Advanced Intelligent Systems and REasoning (ZLAIRE) is taking a leadership role in this journey. ZLAIRE is pioneering the development of advanced intelligent reasoners. Two of its key objectives are to explore the ethical and philosophical implications of AI and develop systems capable of moral reasoning and decision-making.

The task of piecing together this complex puzzle of AI development, standardization, experimentation, and philosophy falls upon the concept of explanation. Explanation acts as a bridge that connects these diverse elements, breaking down complexities, demystifying processes, and helping us to understand both real and imagined worlds.

ZLAIRE's focus will pivot sharply towards harnessing logic for AI reasoning, a step that promises to revolutionize a variety of disciplines, from philosophy to computer science. Our lab is committed to enhancing the reasoning capabilities of these advanced systems, laying the groundwork for a future where AI reasoning will play an increasingly central role in our lives.

Structure of this paper. Section 2 introduces the role of standardization, experimentation and explanation. Section 3 presents examples of intelligent reasoners, such as the Jiminy moral advisor [26], a multiagent deontic argumentation system, and new perspectives on balancing in decision-making and dialogues for moral persuasion. Section 4 concludes the article with some observations on creating a platform for experimental user studies for AI ethics and explainable AI.

2 Standardisation, Experimentation and Explanation

In this section, we discuss methodologies that address three key challenges in advanced intelligent systems and reasoning: standardization, experimentation

and explanation.

2.1 Standardization: Deontic ASP

Answer set programming (ASP) is a prominent paradigm for knowledge representation and reasoning, known for its wide range of applications and efficient tools like `clingo` and `DLV`. ASP's success is attributed to its solid theoretical foundations, including its logical characterization based on equilibrium logic.

Answer set programming plays a crucial role in the standardization of AI reasoners by providing a well-defined and expressive formalism for knowledge representation and reasoning. Its ability to handle complex and nonmonotonic reasoning tasks, along with its solid theoretical foundations based on equilibrium logic, makes ASP an essential candidate for standardization efforts in the AI community. By offering a standardized framework, ASP enables researchers and developers to build interoperable reasoning systems, promotes the sharing and exchange of knowledge representation models, and fosters the development of efficient and powerful reasoning tools. The standardization of AI reasoners through ASP facilitates collaboration, advances the field, and contributes to the broader adoption of AI technologies in various domains.

Deontic logic is commonly combined with nonmonotonic reasoning techniques to represent and reason about norms. Some tools for defeasible deontic logic have been introduced, but standardization and flexibility are still lacking. In a recent paper, Cabalar, Ciabattoni, and Van der Torre [13] presented a deontic extension of equilibrium logic, focusing on reasoning about literals with explicit negation ("classical" negation in ASP). This extension is encoded in ASP while maintaining the same computational complexity.

2.1.1 Logic Programs

We recall the definition of answer sets for propositional logic programs with explicit negation. We start from a propositional *signature*, a set of atoms At, and define an *explicit literal* as any $p \in At$ or its explicit negation $\neg p$. A *default literal* is any explicit literal L or its default negation $\sim L$. A *rule* is an implication of the form:

$$H_1 \vee \cdots \vee H_n \leftarrow B_1 \wedge \cdots \wedge B_m \tag{1}$$

where $n, m \geq 0$ and all H_i and B_j are default literals. The disjunction $H_1 \vee \cdots \vee H_n$ in (1) is called the rule *head*. When $n = 0$, the head is the empty disjunction $\bot$, and the rule is said to be a *constraint*.

The conjunction $B_1 \wedge \cdots \wedge B_m$ in (1) is called the rule *body*. When $m = 0$, it corresponds to the empty conjunction $\top$ and, when this happens, we normally omit both the body $\top$ and the $\leftarrow$ symbol. Moreover, if $m = 0$, $n = 1$, and the head consists of a unique explicit literal H_1 (no default negation), we say that the rule is a *fact*. A *logic program* is a set of rules. For the sake of simplicity, this paper deals with finite programs which we sometimes represent as the conjunction of their rules. Logic programs may contain variables, but they are understood as an abbreviation of all their possible ground instances (for simplicity, we do not allow function symbols).

A *propositional interpretation* T for a signature At is any set of explicit literals that is *consistent*, i.e., it contains no pair of literals p and $\neg p$ for the same atom $p \in At$. Given any rule r like (1) containing no default negation, we say that an interpretation *satisfies* r if there is some head explicit literal $H_i \in T$ whenever all body literals $B_j \in T$. The *reduct* of a logic program Π with respect to an interpretation T, written Π^T, is the result of: (1) removing all rules with a default literal $\sim L$ in the body such that $L \in T$, (2) removing all rules with a default literal $\sim L$ in the head such that $L \notin T$, and (3) removing the rest of the default literals. An interpretation T is an *answer set* of a logic program Π if it is $\subseteq$-minimal among all the interpretations satisfying all the rules of Π^T.

2.1.2 Deontic Logic Programs

Following a minimalist approach, Cabalar et al. [13] extended ASP with two new types of propositions to handle atomic *obligations* Op (read as "p is obligatory") and atomic *prohibitions* Fp ("p is forbidden"), for any atom $p \in At$. In many deontic logics, a prohibition Fp can be defined as an obligation $O\neg p$. However, deontic ASP refrains from reading O and F as real operators, seeing them as prefixes for new ASP atoms called "Op" and "Fp" in the signature. Keeping p, Op and Fp separate as three independent propositions makes sense since, for instance, there is no established connection between Op and p, as one may have the obligation to do p but p may not hold (i.e., the obligation is not fulfilled), and similarly for prohibitions. In addition, under certain conditions, Cabalar et al. [13] allow Op and Fp to hold together.

2.2 Experimentation: LogiKEy

The Logic and Knowledge Engineering Framework and Methodology (LogiKEy) [6,7] offers a framework and methodology for utilizing normative theories and deontic logics to create explicit ethico-legal control and governance mechanisms for intelligent autonomous systems. The formalization results of their ongoing work can be found publicly on the LogiKEy repository at www.logikey.org.

LogiKEy's cohesive formal framework is grounded in shallow semantical embeddings (SSEs) of deontic logics, combinations of logics, and ethico-legal domain theories within an expressive classic higher-order logic (HOL). To corroborate our approach, we have incorporated the primary strands of current deontic logic within HOL, and have been testing this approach for several years.

2.2.1 Three Layers

The methodology of LogiKEy assists logic and knowledge engineers in the concurrent development of three layers: L1 consists of logics and their combinations, L2 is concerned with ethico-legal domain theories, and L3 contains concrete examples and applications.

These three levels are related as follows. Normative governance applications, developed at layer L3, are reliant on ethico-legal domain theories drawn from layer L2. These theories are in turn formalized within a specific logic or logic

combination provided at layer L1.

The engineering process across these layers includes points for backtracking and may require several iterations. Higher layers may also demand modifications to the lower layers. Such potential requests, unlike most other methods, may also involve significant modifications to the logical foundations engineered at layer L1. These changes at the logic layer are flexibly facilitated in our meta-logical approach.

2.2.2 Experimentation

This meta-logical strategy provides robust tool support. Existing theorem provers and model finders for HOL help the LogiKEy designer to create ethically intelligent agents, offering the flexibility to experiment with foundational logics and their combinations, ethico-legal domain theories and specific examples simultaneously. Continuous enhancements of these ready-made provers inadvertently boost reasoning performance within LogiKEy.

The availability of powerful systems like Isabelle/HOL [32] and Leo-III [39] allows us to transform formal ethics along the line of our approach. Although adopting HOL might be a paradigm shift for ethical reasoning, this insight is already well established in formal deduction. While deontic logic representation in HOL isn't straightforward, once achieved, minor changes and their effects become much more manageable. This aligns perfectly with how our approach aids the design of normative theories for ethico-legal reasoning. The ease with which users can modify and adapt existing theories makes the design of normative theories accessible to non-specialist users and developers.

2.3 Explanation: Three Faces of Argumentation

As AI systems increasingly permeate our daily lives, the way in which they explain themselves to and interact with humans becomes an increasingly critical research area. Formal argumentation, as understood in AI, can provide a general, unifying framework for explanations, combining aspects from knowledge representation and reasoning, and human-computer interaction. Formal argumentation has developed into a rich and multidimensional field that encompasses various perspectives and approaches to the study of reasoning, persuasion, and decision-making. In formal argumentation, different branches have emerged. Argumentation as inference includes abstract and structured argumentation (Dung, 1995; Modgil et al., 2014; Toni and Tamma, 2014), offering a systematic framework for analyzing and evaluating arguments, taking into account their logical structure. Argumentation as dialogue (Arisaka et al., 2022) explores multiagent systems and strategic interactions, focusing on the dynamics of various kinds of dialogues. Argumentation as balancing (Gordon and Walton, 2007) addresses the need to strike a balance between conflicting viewpoints and has found applications in domains such as law and ethics.

2.3.1 Argumentation as Inference

Argumentation as inference fosters clarity and systematic understanding of arguments. It helps make reasoning systems capable of formulating coherent

and logical conclusions. One of the strengths of the abstract argumentation framework is its powerful generality. Its process of transforming a knowledge base into an argumentation graph and obtaining a set of acceptable conclusions for that knowledge base has been dubbed "the argumentation pipeline" [23]. In more detail, the argumentation pipeline takes input from a knowledge base in a formal language that specifies how arguments are constructed from a premise set as well as a number of inference rules. Premises are formulas in a given formal language. They represent the evidence or information on which arguments are based. Rules are used to infer new formulas from others. Arguments are thus considered to be the result of applying inference rules to premises and, possibly, chaining such applications. As a second step, attack relations are established between the arguments, taking various considerations about the arguments into account (such as their syntactic form, their strength, and so on). Argumentation semantics are then used to obtain sets of acceptable arguments based on the argumentation graph constructed in the previous step. Finally, sets of acceptable conclusions are obtained on the basis of the sets of acceptable arguments. Such a knowledge base can be used to model, for example, default reasoning [43], logic programming with negation as failure [18], and autoepistemic reasoning [11]. In this regard, one potential future direction for research is causal argumentation [10], particularly due to the limitations of existing rule-based systems in representing causal knowledge. Another critical aspect that requires attention is the identification and exploration of specific argument types associated with causality, such as those incorporating counterfactual statements. There are three central approaches that correspond to this line of research: logic-based deductive methods [8,1,9], assumption-based argumentation systems [11,41], and ASPIC systems [30].

One important development is the study of rationality postulates as introduced by Caminada and Amgoud [14,15] and later extended by Caminada et al. [19] and Wu and Podlaszewski [42]. They proposed several properties that any argumentation system should fulfil. These properties are meant to ensure that argumentation-based inferences make sense from a logical point of view, i.e., that the graph-based selection is sensible from the perspective of the logical language that was used to construct the argument graph. The choice of attack relation (e.g., unrestricted versus restricted rebut) can have a major impact on the satisfaction of the rationality postulate.

2.3.2 Argumentation as Dialogue

Argumentation dialogues, where the role of agents is on the central stage, have been significantly applied to the fields of AI and law and multiagent systems since the 1990s (see Prakken [3, Chap. 2]). In the early days, Lorenzen and Lorenz [28] developed formal dialogue systems for argumentation using a game formulation of disputes among agents. The acceptance of an argument provided by an agent depends on several aspects, such as trust [37,24], and voting in social choice [20,25,2,17]. In 2011, Rienstra et al. [38] proposed multi-sorted argumentation, where each agent owns a part of the framework and may locally

adopt different semantics. Multiagent systems can be roughly grouped into two categories: cooperative and non-cooperative [22]. In cooperative systems, agents share a common goal and fully cooperate to achieve it. Agents can form coalitions to improve their performance, i.e., pooling their efforts and resources to achieve particular tasks at hand more efficiently [21]. In a non-cooperative system, each agent has its own desires and preferences, which may conflict with those of other agents. Multiagent argumentation takes inspiration from several disciplines such as game theory, and it can be further developed towards coalitional game theory by introducing the notion of coalition and associate arguments of (sets of) agents. An alternative approach to multiagent argumentation takes its inspiration from voting theory, and more generally from social choice.

2.3.3 Argumentation as Balancing

In Chapter 3 of the Handbook of Formal Argumentation, Thomas Gordon proposed an alternative definition of argumentation highlighting the importance of argumentation for making justified decisions [3, Chap.3]. Argumentation is thus not only important when resolving conflicts of opinion in persuasion dialogues, but also when deciding courses of action in deliberation dialogues [3, Chap.3]. He then gave a new definition of argumentation: argumentation is a rational process, typically in dialogues, for making and justifying decisions about various kinds of issues. In this application, pro and con arguments provide alternative resolutions of the issues, so that the options (or positions) are put forward, evaluated, resolved and balanced. Argumentation as balancing finds significant applications in the realms of law and ethics. In these domains, the objective is not merely to assess the validity or strength of individual arguments but to strike a balance between conflicting viewpoints or interests. Balancing involves weighing different considerations, evaluating the relative importance of arguments, and reaching decisions that are ethically sound and legally justifiable.

3 Examples of Advanced Intelligent Reasoners

This section reviews some recent examples of advanced intelligent systems and future research lines in this area.

3.1 The Jiminy Moral Advisor

Autonomous agents such as self-driving cars and smart speakers are aware of a range of possible actions they can take in a given situation. As some of these actions might affect people nearby (drivers, passengers, pedestrians and resp. household members), these agents' behavior should adjust to some given moral regulation. Next, we describe our recent work [26] in this research area, based on deontic argumentation.

Machine ethics can be tackled in two different ways [31]. So-called morally implicit agents are provided with contextual rules for their ethical labeling of actions —with only actions labeled as good being permitted. Morally explicit agents, on the other hand, make moral judgments, or are given guidelines or examples they can extrapolate from about good and bad actions.

For a given agent, relevant stakeholders are (types of) human beings potentially affected by that agent. It has been argued that all these types of people should be given a voice in the regulation of this agent [5]. (The alternative is regulation by a single stakeholder, who might be tempted to look after their own particular interests.) A natural way of letting these voices be heard is a normative system. Observe that, in contrast to Section 2.1, no explicit use of obligation or permission modalities is made in the language.

Definition 3.1 A *normative system* of stakeholder s is a tuple $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s)$ where:

- $\mathcal{L}$ is a logical language over a set of atoms *Var*;
- $^-: \mathcal{L} \mapsto 2^{\mathcal{L}}$ is a (partial) contrariness function $\overline{\varphi} = \{\psi_1, \ldots, \psi_k\}$ that extends logical negation $\neg\varphi \in \overline{\varphi}$;
- $\mathcal{R}_s^\tau$ is a set of norms $\phi_1, \ldots, \phi_n \Rightarrow_s^\tau \phi$ where $\tau \in \{r, c, p\}$ denotes a regulative, constitutive and resp. permissive norm; we also write $\mathcal{R}_s = \mathcal{R}_s^r \cup \mathcal{R}_s^c \cup \mathcal{R}_s^p$.

Given a set of facts $\mathcal{K}$, the *argumentation theory* of stakeholder s is the tuple abusively denoted $\mathcal{N}_s = (\mathcal{L}, ^-, \mathcal{R}_s)$. For a set of stakeholders $\mathcal{S} = \{s_1, \ldots, s_n\}$, the argumentation theory is the tuple $\mathcal{N}_\mathcal{S} = (\mathcal{L}, ^-, \mathcal{R}_\mathcal{S}, \mathcal{K})$ defined by $\mathcal{R}_\mathcal{S} = \mathcal{R}_{s_1} \cup \cdots \cup \mathcal{R}_{s_n}$.

Note that elements $\mathcal{L}, ^-, \mathcal{K}$ are shared among all the stakeholders. While $\mathcal{K}$ is a collection of brute facts, institutional facts can be detached from brute facts and constitutive norms in $\mathcal{R}^c$. Institutional facts describe high-level facts (such as legal claims) in the scenario (whether an utterance is a threat, whether a bike counts as a vehicle, etc.).

Example 3.2 A smart speaker scenario involves three stakeholders: $L = law$, $H = human\ users$ and $M = manufacturer$. The norms and facts are:

$$\mathcal{R}_L = \left\{ \begin{array}{c} D \textit{ is made by } M \Rightarrow_L^r M \textit{ is } \textbf{\textit{law}} \textit{ compliant}, \\ M \textit{ is a } \textbf{\textit{business}} \textit{ in Norway} \Rightarrow_L^r \textit{comply with the } \textbf{\textit{GDPR}} \end{array} \right\}$$

$$\mathcal{R}_H = \left\{ \begin{array}{c} D \textit{ collects data} \Rightarrow_H^r \textbf{\textit{protect}} \textit{ privacy}, \\ D \textit{ finds a threat} \Rightarrow_H^r \textbf{\textit{report}} \textit{ threat} \end{array} \right\}$$

$$\mathcal{R}_M = \left\{ \begin{array}{c} D \textit{ finds a threat} \Rightarrow_M^r \textbf{\textit{collect}} \textit{ data w.o. permission}, \\ M \textit{ is registered in Norway} \Rightarrow_M^c M \textit{ is a } \textbf{\textit{business}} \textit{ in Norway} \end{array} \right\}$$

$$\mathcal{K} = \left\{ \begin{array}{c} D \textit{ is made by } M,\ D \textit{ collects data}, \\ D \textit{ finds a threat},\ M \textit{ is registered in Norway} \end{array} \right\}$$

Let $\mathcal{R}_s = \{S_1, S_2\}$ for each stakeholder s. Contrary formulas (omitted here) give rise to the next conflicts between norms, expressed with arrows:

$$L_1 \leftrightharpoons H_1 \qquad H_1 \rightarrow H_2 \qquad H_1 \leftrightharpoons M_1 \qquad M_1 \rightarrow L_1 \qquad L_2 \leftrightharpoons M_1$$

Following [34], a priority relation between rules is designed with moral recommendations in mind. First, deontic detachment (the chaining of regulative norms) is not considered for the detachment of remote obligations. Secondly,

where there is conflict, (1) permission norms are understood as exceptions to (and hence preferred to) regulative norms, and (2) current facts in $\mathcal{K}$ also take preference over regulative norms. Finally, for hard cases, we can endow the Jiminy advisor with a specific set of contextual preferences over stakeholders, the latter judged as better or worse normative sources in particular scenarios.

Definition 3.3 A *priority relation* $\preceq$ is defined as follows: first, it applies both ways between any pair of rules of the same τ-type; secondly, its strict fragment $\prec = \preceq \cap \not\succeq$ applies to regulative rather than permissive or constitutive norms (or facts). In sum, for any stakeholders s, s' and norm type $\tau \neq p$,[1] More precisely, the priority relation consists of the following three sets:

$$\mathcal{R}_s^\tau \times \mathcal{R}_{s'}^\tau \subseteq \preceq \qquad \mathcal{R}_s^r \times \mathcal{R}_{s'}^p \subseteq \prec \qquad \mathcal{R}_s^r \times \mathcal{R}_{s'}^c \subseteq \prec .$$

Two semantics for these normative systems can be given: first in terms of norm extensions, i.e., from consistent sets of norms, and secondly as ASPIC+ style arguments.

- A *norm extension* E is (the heads of) a maximally consistent set of norms built with a priority order for facts and permissions in its construction.
- An *argument extension* $\mathcal{E}$ is a set of arguments defined by one of the common Dung semantics: admissible, complete, preferred, grounded, or stable.

From norm extensions (or argument extensions), one can detach the corresponding obligations using brute or institutional facts. Figure 1 shows the argumentation approach with a schematic illustration of the arguments generated from the norms and facts listed in Example 3.2.

Example 3.4 Continuing with the smart speaker example, the following consistent sets of obligations are detached from two norm extensions E_1, E_2:

$$Obl(E_1) = \{\textbf{\textit{protect}}\ \textit{privacy}, \textit{comply with the}\ \textbf{\textit{GDPR}}\}$$
$$Obl(E_2) = \{\textbf{\textit{report}}\ \textit{threat}, \textbf{\textit{collect}}\ \textit{data w.o. permission}\}.$$

For the argumentation approach, arguments $\{A_1, \ldots, A_8\}$ are generated by combining facts and norms of stakeholders; see also Figures 1 and 2 below.

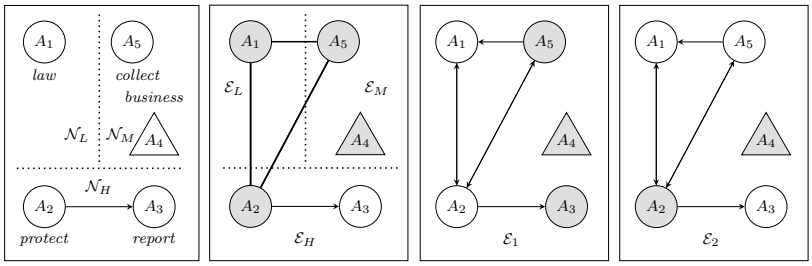
Definition 3.5 Given a collection $\mathcal{C}$ of semantic extensions, a *moral conflict* in $\mathcal{C}$ is a pair of contrary obligations $\varphi \in \overline{\psi}$ within $\mathcal{C}$:

$$\varphi \in Obl(E_1), \psi \in Obl(E_2) \text{ for some } E_1, E_2 \in \mathcal{C}.$$

In argumentation semantics, a more fine-grained distinction of conflicts can be made between direct attacks, where the priority relation $\preceq$ suffices to defeat, and indirect attacks, which requires a strict priority $\succ$ for the attacked argument to become the defeater. The two semantics are related as follows:

i. (complete, preferred, grounded, stable): any argument extension satisfies the rationality postulates [16];

[1] Contrary permissions, say for p and $\neg p$, do not give rise to a deontic conflict. We enforce this property through the absence of a $\preceq$-priority between the corresponding permission rules.



(a) Level 1. Individual frameworks (b) A preferred extension (c) Level 2. Combined framework (d) Level 2. Combined framework

Fig. 1. Obligation and institutional arguments are represented as circles and triangles respectively, and are labeled with their conclusions. (a) individual frameworks for $\mathcal{N}_L, \mathcal{N}_M, \mathcal{N}_H$; (b) the preferred extension (in gray) of each framework; thick lines denote moral dilemmas; (c)–(d) the combined framework (Level 2) with a preferred extension in each subfigure.

ii. (complete, preferred, grounded, stable): any argument extension $\mathcal{E}$ extends into a norm extension E, e.g., $\mathcal{E} \subseteq E$; (stable): for the stable semantics, we moreover have $\mathcal{E} = E$;

iii. (complete, preferred, grounded): under a symmetric contrariness function $^-$, any norm extension E extends some argument extension $\mathcal{E}$, i.e., $E \supseteq \mathcal{E}$;

iv. (naive): the set of norm extensions E corresponds exactly to the set of argument extensions $\mathcal{E}$ under naive semantics.

The Jiminy moral advisor identifies moral dilemmas at four different levels, and proceeds to resolve them by moving to the next level.

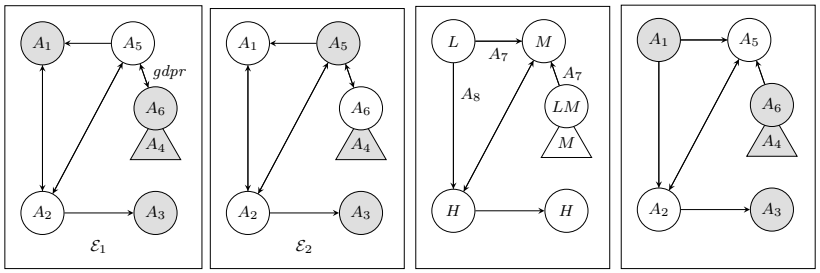
1. **Individual frameworks.** Each stakeholder builds its own argumentation framework using only its own norms.
2. **Combined framework.** All arguments from level 1 are put together.
3. **Integrated framework** All the stakeholders' norms can combine into arguments.
4. **Reduced framework** Jiminy's specific preferences between stakeholders are added. Jiminy arguments can revise the defeat relation.

Figures 1–2 illustrate the four levels and the identification and resolution of moral dilemmas in each level.

Definition 3.6 A Jiminy preference norm is an expression of the form $\varphi_1, \ldots, \varphi_n \Rightarrow s \succ s'$ where $s \neq s'$ are stakeholders. This reads as: *in situations where $\varphi_1, \ldots, \varphi_n$ hold, $\mathcal{R}_s$-norms take priority over $\mathcal{R}_{s'}$-norms.*

Example 3.7 The Jiminy preference norms are the following:

$$\mathcal{R}_J = \left\{ \begin{array}{l} D \textit{ collects data} \Rightarrow L \succ M, \\ D \textit{ finds a threat} \Rightarrow L \succ H, \\ \neg D \textit{ finds a threat} \Rightarrow H \succ L \end{array} \right\}$$



(a) Level 3. Integrated framework (b) Level 3. Integrated framework (c) Level 4. Reduced framework (d) A compatible extension

Fig. 2. (a) the integrated framework (Level 3) with one preferred extension $\mathcal{E}_1$; note the new argument for *gdpr*; (b) another extension $\mathcal{E}_2$; (c) the reduced framework (Level 4) with the introduction of preference norms; a comparison of arrows with (a)–(b) shows how arguments A_7, A_8 revise the defeat relation; (d) extension $\mathcal{E}_1$; it is compatible with the revised defeat, while $\mathcal{E}_2$ is not (not shown).

The defeat relation between two arguments can be revised based on a comparison of the stakeholders' contribution to the norms of each argument.

Levels 1–3 apply Dung semantics as usual over the corresponding argumentation framework. The reduced framework, following Brewka [12], introduces a two-step procedure where: (1) extensions are computed, including arguments expressing Jiminy preferences; the accepted Jiminy arguments revise the original defeat relation, and so (2) one checks if the original extension is still an extension under the new defeat; if so, we say that the extension is *compatible* with the defeat induced. Moral dilemmas are checked in the compatible extensions.

Example 3.8 Three extensions exist for the integrated framework (Level 3), two of which are shown in Figure 2 as $\mathcal{E}_1$ and $\mathcal{E}_2$. At the reduced framework, only one compatible extension remains: $\mathcal{E}_1^+ = \mathcal{E}_1 \cup \{A_7, A_8\}$, and so all moral dilemmas have been resolved at Level 4. The Jiminy returns the obligations:

$$Obl(\mathcal{E}_1^+) = \{M \textit{ is } \textbf{\textit{law}} \textit{ compliant}, \textit{comply with the } \textbf{\textit{GDPR}}, \textbf{\textit{report}} \textit{ threat}\}.$$

In the next section, we discuss several limitations of this centralized approach to multi-agent deontic argumentation. The Jiminy advisor we have just described will be called Autonomous Jiminy from now on.

3.2 Dialogues for Machine Ethics

As described in the previous section, the Autonomous Jiminy (AJ) moral advisor combines norms into arguments, identifies their conflicts as moral dilemmas, and evaluates the arguments to resolve each dilemma (whenever possible). One weakness of this approach is that stakeholders have no control over how their norms will be used to pass a moral recommendation to the agent. A research line to be explored in the future is letting the stakeholders' avatars participate directly in discussions about moral recommendations for the agent. These two approaches illustrate the distinction between argumentation as logic (Sec. 2.3.1)

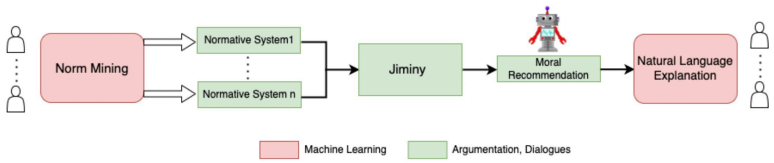


Fig. 3. The Dialogue Jiminy, the agent and the stakeholders.

and argumentation as dialogue (Sec. 2.3.2); see also Prakken [4, Ch. 2].

A Dialogue Jiminy (DJ) for machine ethics (ME) can resolve an agent's moral dilemmas through a persuasion dialogue between the avatars. In contrast to an AJ, a DJ will preserve the stakeholders' autonomy by letting the avatars choose strategies in the dialogue about recommending moral choices. The DJ can also feature a bidirectional language interface to facilitate the normative programming of its avatar and provide it with explanations (see Figure 3).

DJ can thus be seen as a first step in an overarching research programme on ME focusing on avatar dialogues and a natural language interface. Case studies can be used to estimate the effects of endowing agents with the DJ dialogue system. Our approach is to adapt or redesign current theories on persuasion dialogues while applying existing large language models (LLMs) to norm mining and explanation generation in the language interface for stakeholders.[2] The design of DJ involves the integration of two different developments:

Dialogues and avatars. Generalise and transform the Autonomous Jiminy moral advisor into an interactive Dialogue Jiminy by replacing argumentation as inference with argumentation as dialogue. Create a communication language and a protocol for persuasion dialogues on moral dilemmas (see [35]). Study strategic aspects of these dialogues for the participants, as in [40].

Norm mining and explanation synthesis. Create a natural language interface between the Dialogue Jiminy and the stakeholders (or the general public). Use machine learning to construct two language modules: (1) use NLP to transform the stakeholders' informal norms into the avatars' formal rules; (2) use natural language generation to synthesise formal dialogues into explanations (in plain language) of why a particular decision was passed as a moral recommendation to the agent.

One can expect to advance the theory of argumentation-as-dialogue in ethical domains and the practical aspects of argumentation-as-inference. On the practical side, we also aim to improve the state of the art in text mining and explainability in AI (XAI) for norms and decisions through a combination of symbolic AI (Dialogue Jiminy) and sub-symbolic or data-driven methods (LLMs). To this end, on needs to:

[2] We are thankful to Davide Liga for his insights on the natural language interface sketched in the present section.

Speech acts	**Attacks**	**Surrenders**
claim C	why C	concede C
C since R	why D (for some D in R) not C since R'	concede D (D in R) concede C
why C	C since R	retract C
concede C		
retract C		

Fig. 4. Persuasion dialogues consist of speech acts (left), listed together with corresponding moves that attack them or surrender.

- identify the speech acts needed for persuasion in ethical decision-making and contrast them with those studied for legal reasoning (see Figure 4);
- design a protocol for persuasion dialogues [36] for ethical domains that complies with the desiderata for formal inter-agent dialogues [29];
- define the avatars, their normative systems, and possible strategies for them;
- study the properties of dialogues and strategies, in line with [33] and [40].

The overall theory will set the stage for next generation dialogue-based moral advisors which stakeholders can substantially contribute to via their avatars. For the language interface, we envisage three key objectives:

- extract relevant norms from natural language (norm mining),
- convert these norms into a formalized language (norm formalization), and
- explain the DJ's output decision (decision explanation).

We will assess the capacity of both generative LLMs (Generative Pre-training Transformer (GPT) or the like) and non-generative LLMs (Bidirectional Encoder Representations from Transformers (BERT) or the like) to fulfill these tasks.

The explanation for the recommended option will be of the form: *these avatars* $a, \ldots, a'$ *successfully convinced all opponents of arguments* $A, \ldots, Z$*, so they retracted their attacks* $A', \ldots, Z'$. Non-generative methods in turn will be used for norm classification by converting our textual data into vector representations, following the positive results obtained, for example, in [27]. This methodology involves retrieving all crucial normative information from classification tasks by recognizing obligation, permission and constitutive rules. Besides the use of available language models, the project will also employ transfer learning techniques to fine-tune these LLMs on all downstream tasks (mining, classification, generation). Transfer learning will allow us to provide LLMs with annotated data and thus create our own specialized, fine-tuned LLMs. These techniques will thus benefit all the tasks related to the language interface described above.

In summary, we aim to make substantial contributions to formal ethics and AI ethics (with the persuasion dialogues), to agent architectures (with the moral council and language interface), and to XAI and human-computer interactions (with the dialogues, argumentation semantics, and again the language interface).

3.3 Balancing for Stakeholders

We now delve into the compelling application of multi-criteria decision-making (MCDM) within the context of autonomous systems that interact with a wide array of stakeholders, each harboring distinct moral interests. In the swiftly advancing landscape of autonomous systems, exemplified by smart speakers and self-driving cars, these entities are assuming progressively pivotal roles within society. As they navigate diverse environments, the intricate nature of their interactions inevitably exposes them to complex scenarios where their actions may have profound implications for drivers, passengers, pedestrians, and household members. Each of these stakeholders, guided by their distinct ethical values and preferences, contribute to a diverse tapestry of moral interests that demand astute attention.

In addressing these moral dilemmas, an intriguing and fruitful approach is to integrate two fundamental methodologies: balancing pros and cons, and case-based reasoning. By carefully weighing the pros and cons of potential actions, the decision-making process can discern the most optimal course of action that aligns with the varied ethical considerations inherent in the given situation. Moreover, leveraging case-based reasoning empowers autonomous systems to learn from past ethical experiences and apply analogous solutions to novel contexts, providing invaluable guidance when confronted with novel moral quandaries.

Incorporating the balancing of pros and cons fosters holistic evaluation of the ethical landscape, enabling the system to navigate delicate trade-offs and prioritize the wellbeing of diverse stakeholders. By systematically quantifying and assigning weights to different ethical criteria, the agent can achieve an equilibrium between competing interests, thus manifesting a thoughtful and morally defensible approach.

At the same time, case-based reasoning endows the autonomous system with the capacity to draw upon an extensive database of historical ethical cases, each capturing the intricacies of distinct moral dilemmas and their resolutions. Armed with this wealth of ethical knowledge, the system can adapt principles from prior cases to novel situations, thereby exhibiting a more contextually attuned ethical acumen.

To further advance this framework, future research could focus on refining the methodology for balancing pros and cons, potentially incorporating adaptive algorithms to dynamically adjust the weights of ethical criteria based on contextual factors. Additionally, delving into the development of more sophisticated case-based reasoning systems, perhaps integrating machine learning techniques to enhance the identification of relevant past cases, presents an enticing avenue to bolster the ethical decision-making capabilities of autonomous systems.

Combining Morally Implicit and Explicit Approaches. The current research area distinguishes between morally implicit agents, who rely on predefined contextual rules for the ethical labeling of actions, and morally explicit agents, who possess the ability to make moral judgments based on guidelines or examples. We propose to explore a hybrid approach that combines elements

of both methodologies. By using morally implicit rules as a foundation, autonomous agents can ensure compliance with basic ethical norms. However, when confronted with novel or ambiguous situations, agents can utilize morally explicit reasoning to extrapolate from previous experiences and apply moral guidelines to unique contexts. This combination may lead to more nuanced and contextually appropriate moral decisions by the agents.

Incorporating Multi-Stakeholder Normative Systems. As the impact of autonomous agents extends to various stakeholders, it is essential to consider the perspectives and preferences of all relevant human beings potentially affected by these agents. To achieve this, we propose to investigate the integration of multi-stakeholder normative systems. These systems allow stakeholders to contribute to the ethical regulation of the agent by expressing their values, beliefs, and ethical norms. By aggregating and reconciling these diverse viewpoints, the agent can behave so as to consider the interests of all affected parties.

Dynamic Ethical Learning and Adaptation. Finally, to ensure the ongoing ethical competence of autonomous agents, we suggest that methods for dynamic ethical learning and adaptation should be explored. As ethical norms evolve over time and new moral considerations arise, agents should be able to update their knowledge base and reasoning mechanisms. By continuously learning from new ethical cases and integrating emerging ethical guidelines, agents can maintain their relevance and effectiveness in adhering to morally-regulated behavior.

In conclusion, by synergistically embracing balancing pros and cons and case-based reasoning, autonomous systems can effectively tackle moral dilemmas stemming from diverse stakeholder perspectives. The integration of these methodologies not only enables agents to navigate intricate ethical landscapes with adeptness but also exhibits a promising direction for advancing ethically competent autonomous agents that conscientiously engage with the complex ethical dimensions of their actions within society.

4 A Platform for User Experiments

We conclude this paper with some observations about the development of a platform for experimental user studies for AI ethics and explainable AI.

4.1 Architecture

The platform for user experiments comprises a logic engine based on Deontic ASP and a chatbot underpinned by a foundation model.

Interoperability between these two components allows seamless exchange of data, enhancing their collective functionality. The logic engine, with its deontic reasoning capabilities, can parse and process complex logical queries. These results are then communicated effectively to the chatbot, which uses its foundation model to generate user-friendly responses.

In terms of use cases, this system is ideal for situations requiring intricate problem-solving. It could be utilized in customer service, where the logic engine dissects complicated user issues and the chatbot provides easy-to-understand solutions. Or it could be applied in an educational context, helping students to

understand complex theories through interactive dialogue.

For user experience, this amalgamation is beneficial. The deontic ASP-based logic engine's advanced reasoning capabilities combined with the natural language processing power of the chatbot results in a system that solves intricate problems and communicates solutions in an accessible and intuitive manner. This ultimately leads to a more satisfying and enriching user experience.

4.2 AI Ethics and Explainable AI

The experimental platform is designed to further AI ethics and explainable AI. It combines: formal methodologies like deontic ASP to create standard knowledge bases and normative systems, LogiKEy for experimentation, and formal argumentation to ensure explanatory clarity. These tools promote a more profound comprehension of moral decision-making within intelligent systems. The platform's objective is to offer a regulated setting for researchers to investigate and scrutinize the ethical consequences of AI-driven decisions.

Logic engines and foundation models, including chatbots, should be viewed as distinct but interconnected components. The logic engines tackle the intricate task of reasoning about ethics, providing systematic and formalized approaches for encapsulating, interpreting and addressing ethical quandaries. On the other hand, chatbots act as the user-facing interface for this logical reasoning, converting highly formal logical outcomes into easy-to-understand, natural language discussions that users can interact with.

A platform that merges these elements can provide a unique path for AI ethics and explainable AI. In this setup, logic engines like deontic ASP would be utilized to map the ethical problem landscape, resolving conflicting norms and reaching ethically optimal solutions. The chatbots, driven by foundation models, would then convey these decisions and the related reasoning to users in an easily comprehendible format, fostering a more interactive, intuitive, and transparent exploration of AI ethics.

4.3 Application Examples

The platform could be utilized to develop ethical AI frameworks. These frameworks would ensure that AI technologies are integrated into society in a way that maximizes their benefits and minimizes their potential harm.

Within the realm of *social robotics*, the platform could be used to develop intelligent systems that improve human-robot interactions, fostering social connections and enhancing overall quality of life.

The platform could facilitate *computational creativity*, helping to develop AI systems capable of innovative thinking. This could revolutionize industries and expand the limits of human imagination.

Within *healthcare*, the platform could be leveraged to optimize AI implementations, improving patient care, enhancing overall wellbeing, and addressing pressing global health issues.

Finally, the platform could be used to develop *explainable AI systems*. These systems would ensure transparency and accountability in AI decision-making, thereby promoting ethical and responsible AI usage.

References

[1] Arieli, O. and C. Straßer, *Sequent-based logical argumentation*, Argument & Computation **6** (2015), pp. 73–99.

[2] Awad, E., J.-F. Bonnefon, M. Caminada, T. W. Malone and I. Rahwan, *Experimental assessment of aggregation principles in argumentation-enabled collective intelligence*, ACM Transactions on Internet Technology (TOIT) **17** (2017), pp. 1–21.

[3] Baroni, P., D. Gabbay and M. Giacomin, "Handbook of Formal Argumentation," College Publications, 2018.

[4] Baroni, P., D. Gabbay, M. Giacomin and L. van der Torre, editors, **1**, College Publications, 2018.

[5] Baum, S. D., *Social choice ethics in artificial intelligence*, AI Soc. **35** (2020), pp. 165–176.

[6] Benzmüller, C., X. Parent and L. van der Torre, *Designing normative theories for ethical and legal reasoning: Logikey framework, methodology, and tool support*, Artificial intelligence **287** (2020), p. 103348.

[7] Benzmüller, C., A. Farjami, D. Fuenmayor, P. Meder, X. Parent, A. Steen, L. van der Torre and V. Zahoransky, *Logikey workbench: Deontic logics, logic combinations and expressive ethical and legal reasoning (isabelle/hol dataset)*, Data in Brief **33** (2020), p. 106409.

[8] Besnard, P. and A. Hunter, *A logic-based theory of deductive arguments*, Artificial Intelligence **128** (2001), pp. 203–235.

[9] Besnard, P. and A. Hunter, *A review of argumentation based on deductive arguments*, Handbook of Formal Argumentation **1** (2018), pp. 437–484.

[10] Bochman, A., *Propositional argumentation and causal reasoning*, , **19**, LAWRENCE ERLBAUM ASSOCIATES LTD, 2005, p. 388.

[11] Bondarenko, A., P. M. Dung, R. A. Kowalski and F. Toni, *An abstract, argumentation-theoretic approach to default reasoning*, Artificial intelligence **93** (1997), pp. 63–101.

[12] Brewka, G., *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.*, 1994, pp. 940–945.

[13] Cabalar, P., A. Ciabattoni and L. van der Torre, *Deontic equilibrium logic with explicit negation*, in: *18th European Conference on Logics in Artificial Intelligence, Dresden, Germany, September 20-22 2023* (2023), pp. 2742–2749.

[14] Caminada, M. and L. Amgoud, *An axiomatic account of formal argumentation*, , **6**, 2005, pp. 608–613.

[15] Caminada, M. and L. Amgoud, *On the evaluation of argumentation formalisms*, Artificial Intelligence **171** (2007), pp. 286–310.

[16] Caminada, M. and L. Amgoud, *On the evaluation of argumentation formalisms*, Artif. Intell. **171** (2007), pp. 286–310.

[17] Caminada, M. and G. Pigozzi, *On judgment aggregation in abstract argumentation*, Autonomous Agents and Multi-Agent Systems **22** (2011), pp. 64–102.

[18] Caminada, M., S. Sá, J. Alcântara and W. Dvořák, *On the equivalence between logic programming semantics and argumentation semantics*, International Journal of Approximate Reasoning **58** (2015), pp. 87–111.

[19] Caminada, M. W., W. A. Carnielli and P. E. Dunne, *Semi-stable semantics*, Journal of Logic and Computation **22** (2012), pp. 1207–1254.

[20] Coste-Marquis, S., C. Devred, S. Konieczny, M.-C. Lagasquie-Schiex and P. Marquis, *On the merging of dung's argumentation systems*, Artificial Intelligence **171** (2007), pp. 730–753.

[21] Elkind, E., T. Rahwan and N. R. Jennings, *Computational coalition formation*, Multiagent systems (2013), pp. 329–380.

[22] Elkind, E., T. Rahwan and N. R. Jennings, *Game theoretic foundations of multiagent systems*, Multiagent systems (2013), pp. 811–848.

[23] Heyninck, J., "Investigations into the logical foundations of defeasible reasoning: an argumentative perspective." Ph.D. thesis, Ruhr University Bochum, Germany (2019).

[24] Huynh, T. D., N. R. Jennings and N. R. Shadbolt, *An integrated trust and reputation model for open multi-agent systems*, Autonomous Agents and Multi-Agent Systems **13** (2006), pp. 119–154.
[25] Leite, J. and J. G. Martins, *Social abstract argumentation*, in: T. Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011* (2011), pp. 2287–2292.
URL `http://ijcai.org/Proceedings/11/Papers/381.pdf`
[26] Liao, B., P. Pardo, M. Slavkovik and L. van der Torre, *The jiminy advisor: Moral agreements among stakeholders based on norms and argumentation*, Journal of Artificial Intelligence Research **77** (2023), pp. 737–792.
[27] Liga, D. and M. Palmirani, *Deontic sentence classification using tree kernel classifiers*, in: K. Arai, editor, *Intelligent Systems and Applications - Proceedings of the 2022 Intelligent Systems Conference, IntelliSys 2022, Amsterdam, The Netherlands, 1-2 September, 2022, Volume 1*, Lecture Notes in Networks and Systems **542** (2022), pp. 54–73.
[28] Lorenzen, P. and K. Lorenz, "Dialogische logik," Wissenschaftliche Buchgesellschaft, 1978.
[29] McBurney, P., S. Parsons and M. J. Wooldridge, *Desiderata for agent argumentation protocols*, in: *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002, July 15-19, 2002, Bologna, Italy, Proceedings* (2002), pp. 402–409.
[30] Modgil, S. and H. Prakken, *The aspic+ framework for structured argumentation: a tutorial*, Argument & Computation **5** (2014), pp. 31–62.
[31] Moor, J. H., *The nature, importance, and difficulty of machine ethics*, IEEE Intelligent Systems **21** (2006), pp. 18–21.
[32] Nipkow, T., L. C. Paulson and M. Wenzel, "Isabelle/HOL - A Proof Assistant for Higher-Order Logic," Lecture Notes in Computer Science **2283**, Springer, 2002.
URL `https://doi.org/10.1007/3-540-45949-9`
[33] Pardo, P. and L. Godo, *A temporal argumentation approach to cooperative planning using dialogues*, in: J. Leite, T. C. Son, P. Torroni, L. van der Torre and S. Woltran, editors, *Computational Logic in Multi-Agent Systems - 14th International Workshop, CLIMA XIV, Corunna, Spain, September 16-18, 2013. Proceedings*, Lecture Notes in Computer Science **8143** (2013), pp. 307–324.
[34] Pigozzi, G. and L. van der Torre, *Arguing about constitutive and regulative norms*, Journal of Applied Non-Classical Logics **28** (2018), pp. 189–217.
[35] Prakken, H., *Formal systems for persuasion dialogue*, Knowl. Eng. Rev. **21** (2006), pp. 163–188.
[36] Prakken, H. and G. Sartor, *Presumptions and burdens of proof*, in: T. M. van Engers, editor, *Legal Knowledge and Information Systems - JURIX 2006: The Nineteenth Annual Conference on Legal Knowledge and Information Systems, Paris, France, 7-9 December 2006*, Frontiers in Artificial Intelligence and Applications **152** (2006), pp. 21–30.
[37] Ramchurn, S. D., D. Huynh and N. R. Jennings, *Trust in multi-agent systems*, The knowledge engineering review **19** (2004), pp. 1–25.
[38] Rienstra, T., A. Perotti, S. Villata, D. M. Gabbay and L. van der Torre, *Multi-sorted argumentation*, in: *International Workshop on Theorie and Applications of Formal Argumentation*, Springer, 2011, pp. 215–231.
[39] Steen, A. and C. Benzmüller, *Extensional higher-order paramodulation in leo-iii*, J. Autom. Reason. **65** (2021), pp. 775–807.
URL `https://doi.org/10.1007/s10817-021-09588-x`
[40] Thimm, M., *Strategic argumentation in multi-agent systems*, Künstliche Intell. **28** (2014), pp. 159–168.
[41] Toni, F., *A tutorial on assumption-based argumentation*, Argument & Computation **5** (2014), pp. 89–117.
[42] Wu, Y. and M. Podlaszewski, *Implementing crash-resistance and non-interference in logic-based argumentation*, Journal of Logic and Computation **25** (2015), pp. 303–333.
[43] Young, A. P., S. Modgil and O. Rodrigues, *Prioritised default logic as rational argumentation*, in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, 2016, pp. 626–634.