

Towards Deontic Explanations Through Dialogue

Kees van Berkel¹, Christian Straßer²

¹Institute for Logic and Computation, TU Wien, Austria

²Institute of Philosophy II, Ruhr University Bochum, Germany

Abstract

Deontic explanations answer why-questions concerning agents' obligations and permissions. Normative systems are notoriously conflict sensitive, making contrastive explanations pressing: "Why am I obliged to do φ , despite my (seemingly) conflicting obligation to do ψ ?" In this paper, we develop a model of contrastive explanatory dialogues for the well-established defeasible reasoning formalism Input/Output logic. Our model distinguishes between successful, semi-successful, and unsuccessful deontic dialogues. We prove that the credulous and skeptical (under shared reasons) entailment relation of Input/Output logic, can be characterized in formal argumentation using preferred and grounded semantics. This result allows us to leverage known results for dialogue models of the latter two semantics. Since this work is the first of its kind, we discuss 5 key challenges for deontic explanations through dialogue.

Keywords

Defeasible normative reasoning, Contrastive deontic explanations, Logical argumentation, Dialogues

1. Introduction

Norms are indispensable in many aspects of society, ranging from law, ethics, to business protocols and AI. They motivate, guide, and regulate agents, whether they are human or artificial. Often, agents affected by norms do not only need to know *that* they are bound by obligations or *that* they may appeal to rights: they need to understand *why*. Such understanding may enhance compliance and collaboration and is especially pressing when conflicts between norms arise. For instance, I may want to know *why I may take over on the left, despite being obliged to drive on the right*. Here, a good explanation not only explains that I am permitted, but also why the obligation to the contrary does not currently apply: the permission is an exception to the obligation. Answers to this type of why-question are called *deontic explanations*.

Deontic logic is the well-established field exploring formal methods to model normative reasoning. However, the focus has been nearly exclusively on formal systems that determine *which* obligations and permissions *can be* inferred from a normative system, rather than to explain *why*. This gap is remarkable, especially given the increasingly vital role that normative systems play in alignment and compliance requirements for AI. This paper investigates how knowledge representation methods can be used to generate explanatory deontic dialogues.


The demand for explanatory models in AI is increasing [1] and formal argumentation provides a promising method in this respect. First of all, formal argumentation has proven to be a unifying framework for nonmonotonic reasoning [2]. In particular, two central paradigms of defeasible

ArgXAI-24: 2nd International Workshop on Argumentation for eXplainable AI

✉ kees.van.berkel@tuwien.ac.at (K. v. Berkel); christian.strasser@rub.de (C. Straßer)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

reasoning, constrained Input/Output (I/O) logic [3] and default logic [4], can be argumentatively characterized [5, 6]. Second, a wide variety of methods has been proposed in Argumentation for Explainable AI (ArgXAI, [7]). Finally, dialogue models and argumentation games [8, 9, 10], offer dynamic characterizations of formal argumentation, that have the potential to yield interactive (or even tailor-made) explanatory episodes through dialogues.

Once a given nonmonotonic logic is represented in logical argumentation, such as I/O- or default logic, dialogical methods can be leveraged for explanatory purposes. However, a first obstacle, in this respect, is that most characterization results are shown with respect to stable semantics (including [5, 6]; also see [2]), whereas other semantics such as preferred, admissible, and grounded are more suitable for dialogical generalization. In brief, the problem with stable extensions is that they reference the entire set of arguments (each argument is either ‘in’ or ‘out’), while we expect explanatory dialogues to focus on reasons relevant to the explanatory purpose. Furthermore, defining dialogue models and argumentation games for skeptical reasoning is challenging (in the context of multi-extension semantics such as preferred; cf. [10]).

Contributions. We provide dialogue models for one of the central defeasible normative reasoning formalisms in the literature: Input/Output logic [3, 11]. Unfortunately, the original formalism does not naturally lend itself to explanatory reasoning. Recently, a highly modular rule-based proof system – the Deontic Argumentation Calculus (DAC) – was developed with the aim of making I/O suitable for explanatory purposes and it was shown that DAC-induced argumentation frameworks are sound and complete for a large class of constrained I/O logics [5] and default logic [6]. Despite these promising results, the correspondences were only obtained for stable semantics, making them seemingly unsuitable for dialogical deontic explanations.

In this article, we extend these results by a model of dialogue episodes for deontic explanations:

- (1) As a preparatory step, we first prove that for DAC-induced argumentation frameworks the stable and the preferred semantics coincide. This allows us to use well-developed preferred dialogue models in the context of I/O reasoning.
- (2) Furthermore, we lift recent results [12] that show that the ‘free consequences’ of skeptical entailment under stable semantics is identical to entailment under the grounded semantics. In other words, we may also use grounded dialogue models for I/O reasoning.
- (3) Using (1) and (2), we enhance dialogue models and define contrastive deontic explanations that explain certain obligations in contrast to seeming obligations to the contrary.

Outline. Section 2 introduces the DAC formalism. In Section 3, we define DAC-induced argumentation frameworks and prove that stable equals preferred and the free consequences correspond to grounded entailment. We harness these results to specify contrastive dialogue models in Section 4. This paper lays the foundations for a more extensive study of dialogue models of deontic explanation and, for this reason, we discuss five key challenges in Section 5.

2. Preliminaries: A Deontic Argumentation Calculus (DAC)

We recall the basics of the Deontic Argumentation Calculus (DAC). Although the results in this paper hold for a range of languages, base logics, and DAC systems, for readability we assume a propositional language \mathcal{L} and classical logic L , and illustrate our approach for one DAC system from [5]. To enhance explainability, \mathcal{L} is labeled and augmented with a language of norms:

Labeled propositional languages: $\mathcal{L}^i = \{\varphi^i \mid \varphi \in \mathcal{L}\}$ where $i \in \{f, o, c\}$.

Norm languages: $\mathcal{L}^n = \{(\varphi, \psi) \mid \varphi, \psi \in \mathcal{L}\}$ and $\overline{\mathcal{L}^n} = \{\neg\Delta \mid \emptyset \subset \Delta \subseteq \mathcal{L}^n, \Delta \text{ is finite}\}$.

Normative systems: $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ is a *normative system*, where $\mathcal{F} \subseteq \mathcal{L}^f$ is a *factual context*, $\mathcal{N} \subseteq \mathcal{L}^n$ a *normative code*, and $\mathcal{C} \subseteq \mathcal{L}^c$ a set of *constraints* (and \mathcal{F} and \mathcal{C} are L-consistent).

Labels explicate the roles that propositional formulas adopt in the reasoning process: φ^f denotes that φ is a fact, φ^o that φ is obligatory, and φ^c that obligations must be consistent with φ . We take $(\varphi, \psi) \in \mathcal{L}^n$ to express the norm “given φ , it is obligatory that ψ ” and $\neg\Delta \in \overline{\mathcal{L}^n}$ is read as “the norms in Δ are jointly inapplicable.” For $\neg\{(\varphi, \psi)\}$, we simply write $\neg(\varphi, \psi)$. The latter type of expression plays an essential role in defeasible reasoning with norms. The entire enhanced I/O language is defined as the union $\mathcal{L}^{io} = \mathcal{L}^f \cup \mathcal{L}^o \cup \mathcal{L}^c \cup \mathcal{L}^n \cup \overline{\mathcal{L}^n}$. We write $\Gamma^i, \Delta^i, \dots$ for finite sets of i -labeled formulas, where $i \in \{f, o, c\}$. We write Γ, Δ, \dots for any finite subset of \mathcal{L}^{io} and Δ^\downarrow for a set $\Delta \subseteq \mathcal{L}^i$ stripped from its label $i \in \{f, o, c\}$.

Defeasible normative reasoning occurs with respect to a normative system \mathcal{K} . The basic idea of I/O reasoning [3] and DAC is that facts (input) trigger norms from which obligations (output) are detached where the constraints filter the output to ensure consistency. Our aim is to construct *arguments* from \mathcal{K} . Our approach belongs to logical argumentation (a subfield of structured argumentation [2]). We write arguments as *sequents*: $a = \Gamma \Rightarrow \varphi$, where $\text{prem}(a) = \Gamma$ is a (possibly empty) set of premises, and $\text{conc}(a) = \varphi$ is the conclusion of the argument. An *explanatory argument* is an argument *stating reasons* for a conclusion. We take facts, constraints, and norms as reasons and differentiate two types of argument:

$$\varphi^f, (\varphi, \psi) \Rightarrow \psi^o \quad \text{and} \quad \varphi^f, \neg\psi^c \Rightarrow \neg(\varphi, \psi)$$

The first type (left) contains arguments providing reasons for obligations, where the fact φ^f and norm (φ, ψ) provide reasons for the obligation ψ^o . The second type (right) contains arguments that attack reasons, expressing which norms are *inapplicable* in the given context, where given φ^f , the norm (φ, ψ) is inapplicable since its detachable obligation is inconsistent with the constraint $\neg\psi^c$. The latter type attacks all arguments using (φ, ψ) as a reason.

A DAC is a sequent-style, that is, rule-based proof system for deriving these two types of argument [5]. We assume that LC is the sound and complete sequent calculus for L.

Deontic Argumentation Calculus (DAC): Let DAC be a system consisting of the rules **Ax**, **FDet**, **DDet**, **Con**, **Ina**, **InaC**, **Taut**, and **Cut** (Figure 1). A DAC-derivation of $\Gamma \Rightarrow \Delta$ is a tree-like structure whose leaves are initial sequents, whose root is $\Gamma \Rightarrow \Delta$, and whose rule-applications are instances of the rules of DAC. We say $\Gamma \Rightarrow \Delta$ is DAC-derivable (written $\vdash_{\text{DAC}} \Gamma \Rightarrow \Delta$) whenever there exists a DAC-derivation for it, $\Gamma \subseteq \mathcal{L}^{io}$, and $\Delta \subseteq \mathcal{L}^{io}$ contains at most one formula. We say $\Gamma \Rightarrow \Delta$ is \mathcal{K} -based whenever $\Gamma \subseteq \mathcal{F} \cup \mathcal{N} \cup \mathcal{C}$.

There are three initial sequent rules: **Ax** introduces labeled versions of any classically derivable $\Gamma \Rightarrow \Delta$ to a DAC-derivation (and so LC rules are not part of DAC). **Taut** guarantees that all propositional tautologies are among the output. **FDet** expresses *factual detachment* and gives an initial explanatory argument stating that the fact φ^f and the norm (φ, ψ) are reasons for concluding the obligation ψ^o . **DDet** corresponds to *deontic detachment* and makes it possible

$$\begin{array}{c}
\frac{\vdash_{\mathcal{L}} \Gamma \Rightarrow \Delta}{\Gamma^i \Rightarrow \Delta^i} \mathbf{Ax}, i \in \{f, o, c\} \text{ and } \Gamma, \Delta \subseteq \mathcal{L} \quad \frac{}{\Rightarrow (\top, \top)} \mathbf{Taut} \quad \frac{}{\varphi^f, (\varphi, \psi) \Rightarrow \psi^o} \mathbf{FDet} \\
\frac{\varphi^f, \Gamma \Rightarrow \Delta}{\varphi^o, \Gamma \Rightarrow \Delta} \mathbf{DDet}^a \quad \frac{\Gamma \Rightarrow \varphi^o}{\Gamma, (\neg\varphi)^c \Rightarrow} \mathbf{Con} \quad \frac{\Gamma, (\varphi, \psi) \Rightarrow}{\Gamma \Rightarrow \neg(\varphi, \psi)} \mathbf{Ina} \\
\frac{\Gamma \Rightarrow}{\Gamma \setminus \mathcal{L}^n \Rightarrow \neg(\Gamma \cap \mathcal{L}^n)} \mathbf{InaC} \quad \frac{\Gamma \Rightarrow \varphi \quad \varphi, \Gamma' \Rightarrow \Delta}{\Gamma, \Gamma' \Rightarrow \Delta} \mathbf{Cut}^b
\end{array}$$

Figure 1: A Deontic Argumentation Calculus (DAC). The upper row represents initial sequent rules. Side-condition (a) on **DDet** stipulates $\Gamma \cap \mathcal{L}^n \neq \emptyset$; and (b) on **Cut** requires that $\varphi \in \mathcal{L}^{io}$.

that a norm may be triggered by obligations detached from other norms (see Ex. 1). The rules **Con**, **Ina**, and **InaC** deal with the defeasibility of normative reasoning and yield attacking arguments. The **Con** rule expresses the consistency constraint that if Γ constitutes reasons for φ^o , then Γ is inconsistent with the constraint $\neg\varphi^c$ (where an empty right-hand side denotes inconsistent reasons). We also refer to $\Gamma \Rightarrow$ as an inconsistent argument. When an argument expresses inconsistent reasons, at least one of its involved norms is inapplicable (**Ina**) and all involved norms are jointly inapplicable (**InaC**). We refer to [5] for other DAC systems.

Example 1. We look at Chisholm’s scenario [11], an archetype of contrary-to-duty reasoning. Billie is obligated to go and help her neighbors (\top, h) (\top denotes that h is detached by default). If Billie goes to help, she must tell the neighbors she goes (h, t) , otherwise she ought not to tell them she goes $(\neg h, \neg t)$. Suppose that Billie does not go and help $\neg h^f$ and, so, violates the default duty in (\top, h) . To know what Billie must do in light of her violation $\neg h^f$, the constraint is imposed that the obligations must be consistent with the fact that Billie does not help $\neg h^c$ [5, 11]. Let $\mathcal{F} = \{\neg h^f\}$, $\mathcal{N} = \{(\top, h), (h, t), (\neg h, \neg t)\}$, and $\mathcal{C} = \{\neg h^c\}$ be the normative system \mathcal{K} . The desired outcome is that Billie ought not to tell the neighbors she goes $\neg t^o$ given that she does not go.

Argument d (below left), stating that Billie ought to tell, is derived with deontic detachment. Argument e (below right), expresses the inapplicability of $\neg(\top, h)$ given the set constraint. Similar reasoning gives the inconsistent argument $x = \neg h^f, (\top, h), (h, t), (\neg h, \neg t) \Rightarrow$, which with **Con**, **Cut**, and **InaC** derives the unattackable $x' = \neg h^f \Rightarrow \neg\{(\top, h), (h, t), (\neg h, \neg t)\}$.

$$\begin{array}{c}
\frac{}{\top^f, (\top, h) \Rightarrow h^o} \mathbf{FDet} \quad \frac{\frac{h^f, (h, t) \Rightarrow t^o}{h^o, (h, t) \Rightarrow t^o} \mathbf{DDet}}{\top^f, (\top, h), (h, t) \Rightarrow t^o} \mathbf{Cut} \quad \frac{\frac{\top^f, (\top, h) \Rightarrow h^o}{\top^f, (\neg h)^c, (\top, h) \Rightarrow} \mathbf{FDet}}{\top^f, (\neg h)^c, (\top, h) \Rightarrow} \mathbf{Con} \\
\frac{}{e = \top^f, (\neg h)^c \Rightarrow \neg(\top, h)} \mathbf{Ina}
\end{array}$$

For the sake of completion, we recall the I/O system out_3 here and some known results [5].

Proposition 1 ([5]). Let $\mathcal{K}^\downarrow = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ be \mathcal{K} stripped from its labels. Let $\Delta \subseteq \mathcal{L}$, $Cn(\Delta) = \{\varphi \mid \Delta \vdash_{\mathcal{L}} \varphi\}$, and $\text{out}(\mathcal{N}, \Delta) = Cn(\{\varphi \mid (\psi, \varphi) \in \mathcal{N} \text{ and } \psi \in Cn(\Delta)\})$. Let $\text{out}_3(\mathcal{N}, \mathcal{F}) = \bigcup_{i \geq 0} O_i$, where $O_0 = \text{out}(\mathcal{N}, \mathcal{F})$ and $O_{i+1} = Cn(O_i \cup \text{out}(\mathcal{N}, O_i \cup \mathcal{F}))$. In words, out_3 is a closure of \mathcal{N} under successive (deontic) detachment with respect to \mathcal{F} . Then $\mathcal{N} \subseteq \mathcal{L}^n$ is \mathcal{C} -consistent in \mathcal{K} , if $\perp \notin Cn(\text{out}_3(\mathcal{N}, \mathcal{F}) \cup \mathcal{C})$. Let $\Theta \subseteq \mathcal{N} \subseteq \mathcal{L}^n$, $\Delta \subseteq \mathcal{F} \subseteq \mathcal{L}$ and $\Omega \subseteq \mathcal{C} \subseteq \mathcal{L}$, we have:

1. $\varphi \in \text{out}_3(\Theta, \Delta)$ iff $\vdash_{\text{DAC}} \Theta, \Delta^f \Rightarrow \varphi^o$;
2. Θ is \mathcal{C} -inconsistent iff there are $\Delta \subseteq \mathcal{F}$ and $\Omega \subseteq \mathcal{C}$ for which $\vdash_{\text{DAC}} \Theta, \Delta^f, \Omega^c \Rightarrow$;
3. $\perp \in Cn(\text{out}_3(\Theta, \Delta) \cup \Omega)$ iff for all $(\varphi, \psi) \in \Theta$, $\vdash_{\text{DAC}} \Theta \setminus \{(\varphi, \psi)\}, \Delta^f, \Omega^c \Rightarrow \neg(\varphi, \psi)$.

3. Formal Argumentation with DAC-arguments

We use formal argumentation [13] to capture the defeasibility of normative reasoning and to explicate norm conflicts [5]. An Argumentation Framework [13] contains a set of arguments and an attack relation between arguments, where semantics stipulate conditions under which sets of arguments are jointly acceptable. We instantiate such frameworks with DAC-arguments.

DAC-induced Argumentation Frameworks Let $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ be a normative system. A DAC-induced argumentation framework $\mathcal{AF}(\mathcal{K}) = \langle \text{Arg}, \text{Att} \rangle$ is defined as follows:

- $\Delta \Rightarrow \Gamma \in \text{Arg}$ iff $\Delta \Rightarrow \Gamma$ is DAC-derivable and \mathcal{K} -based.
- a defeats b , i.e., $(a, b) \in \text{Att} \subseteq \text{Arg} \times \text{Arg}$ iff $\text{conc}(a) = \neg \Delta \in \overline{\mathcal{L}^n}$, and $\Delta \subseteq \text{prem}(b)$.

We write $\text{Arg}(\Sigma) = \{\vdash_{\text{DAC}} a \mid \text{prem}(a) \subseteq \Sigma\}$.

Argumentative Semantics and Entailment Let $\langle \text{Arg}, \text{Att} \rangle$ be an \mathcal{AF} and let $\mathcal{S} \subseteq \text{Arg}$: \mathcal{S} defeats an argument $a \in \text{Arg}$ if there is a $b \in \mathcal{S}$ that defeats a ; and \mathcal{S} defends a if \mathcal{S} defeats every argument that defeats a . Let $\text{Defended}(\mathcal{S})$ be the set of arguments defended by \mathcal{S} .

We recall the following semantic definitions [13]: \mathcal{S} is *conflict-free* if it does not defeat any of its own elements; \mathcal{S} is *admissible* if it is conflict-free and defends all $b \in \mathcal{S}$; \mathcal{S} is *preferred* if it is maximally admissible; \mathcal{S} is *stable* if it is conflict-free and defeats all $b \in \text{Arg} \setminus \mathcal{S}$; \mathcal{S} is *grounded* if $\mathcal{S} = \bigcup_{i \geq 0} \mathcal{G}_i$ where $\mathcal{G}_0 = \emptyset$ and $\mathcal{G}_{i+1} = \text{Defended}(\mathcal{G}_i)$.

Let $\text{sem} \in \{\text{admissible, preferred, stable, grounded}\}$, we define two *entailment relations*:

- $\mathcal{AF} \vdash_{\text{sem}}^{\cap \text{rea}} \varphi$ iff there is an a contained in every sem-extension that concludes φ ;
- $\mathcal{AF} \vdash_{\text{sem}}^{\cup} \varphi$ iff there is a sem-extension \mathcal{E} for which there is an $a \in \mathcal{E}$ concluding φ .

$\vdash_{\text{sem}}^{\cap \text{rea}}$ captures the shared arguments (*shared reasons*) by all sem-extensions. The resulting conclusions are called the *free consequences* of \mathcal{K} , which are obligations from unproblematic norms compatible with any sem-extension. Credulous entailment $\vdash_{\text{sem}}^{\cup}$ captures the existence of reasons in favor of a conclusion for some sem-extension, expressing a *defensible stance*.

Example 2. The partial \mathcal{AF} in Figure 2 captures the scenario from Ex. 1. There is only one stable extension $\{a, b, g, x'\}$ (the arrows from b, e, f , and g to x are implicit), which is also the grounded extension (cf. Prop. 3-2 below). We may, thus, conclude $\mathcal{AF} \vdash_{\text{sem}}^{\cap \text{rea}} (\neg t)^o$ (where $\vdash \in \{\vdash_{\text{sem}}^{\cap \text{rea}} \mid \star \in \{\cap \text{rea}, \cup\}\}$ and $\text{sem} \in \{\text{stable, grounded}\}$). As desired, since Billie does not go to help her neighbors, she ought not to tell them she is coming. Billie may now ask “Why am I obliged to not tell my neighbors, despite my seeming duty to tell them I am coming to help?” To this we turn next.

We recall [5, Theorem 2] that for the system adopted in this paper, DAC-induced \mathcal{AF} s are sound and complete for the system out_3 of constrained Input/Output logic [3].

Proposition 2. Let $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ and let $\text{maxfam}(\mathcal{K}) = \{\mathcal{N}' \subseteq \mathcal{N} \mid \perp \notin \text{Cn}(\text{out}_3(\mathcal{N}', \mathcal{F}^\downarrow) \cup \mathcal{C}^\downarrow)\}$ and for each $\mathcal{N}' \subseteq \mathcal{N}$, $\perp \in \text{Cn}(\text{out}_3(\mathcal{N}', \mathcal{F}^\downarrow) \cup \mathcal{C}^\downarrow)\}$ be the set of maximal consistent sets of norms over \mathcal{K} . Let \mathcal{AF} be induced by DAC and \mathcal{K} with the set of stable extensions $\text{stable}(\mathcal{AF})$:

1. If $\mathcal{N}' \in \text{maxfam}(\mathcal{K})$, then $\text{Arg}(\mathcal{F} \cup \mathcal{N}' \cup \mathcal{C}) \in \text{stable}(\mathcal{AF})$;
2. If $\mathcal{A} \in \text{stable}(\mathcal{AF})$, then there is a $\mathcal{N}' \in \text{maxfam}(\mathcal{K})$ for which $\mathcal{A} = \text{Arg}(\mathcal{F} \cup \mathcal{N}' \cup \mathcal{C})$.

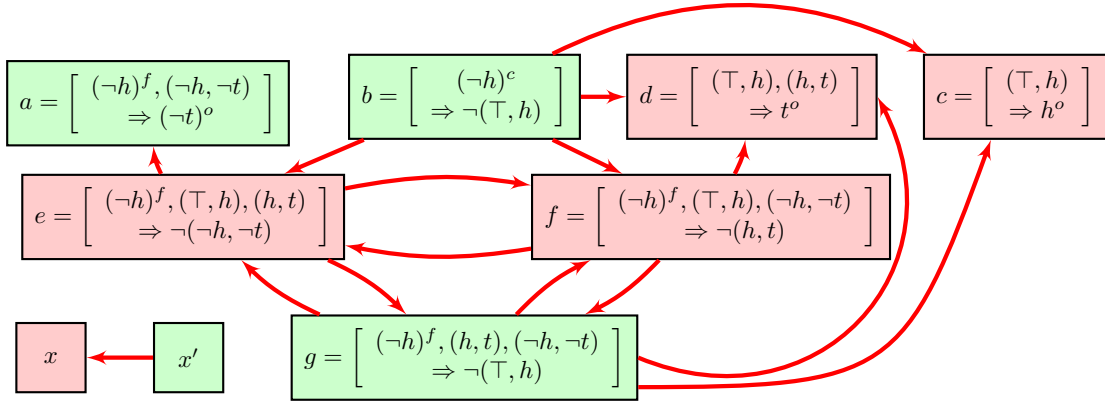


Figure 2: DAC-induced \mathcal{AF} of Ex. 1. Arrows denote defeats relative to the constraint $\mathcal{C} = \{(-h)^c\}$.

Our aim is to employ dialogue models for contrastive deontic explanations and for this we need some additional results. Since the grounded extension is unique [13], $\vdash_{\text{grounded}}^{\text{rea}}$ and $\vdash_{\text{grounded}}^{\text{u}}$ coincide and we simply write \vdash_{grounded} . The proofs below do not reference specific DAC-rules (outside the base system [5, 12]), and thus generalize to all DAC systems in [5] (augmented with **InaC**). Proposition 3 tells us that for reasoning about the free consequences under the stable semantics, it suffices to reason with the grounded semantics. Proposition 4 shows that the preferred and stable semantics coincide for DAC. Consequently, these propositions allow us to apply well-developed dialogue techniques to DAC for skeptical (in terms of free consequences) and credulous reasoning under the grounded, respectively the preferred semantics.

Proposition 3. *Let \mathcal{AF} be a DAC-induced \mathcal{AF} for $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ and let $\text{stb}(\mathcal{AF})$ and $\text{grd}(\mathcal{AF})$ be the set of stable extensions, respectively the grounded extension of \mathcal{AF} :*

1. $a \in \bigcap \text{stb}(\mathcal{AF})$ iff every defeater $b \in \text{Arg}$ of a is \mathcal{C} -inconsistent (i.e., it is defeated by an argument $c \in \text{Arg}(\mathcal{F} \cup \mathcal{C})$).
2. $\text{grd}(\mathcal{AF}) = \bigcap \text{stb}(\mathcal{AF}) = \mathcal{G}_2 = \text{Defended}(\text{Arg}(\mathcal{F} \cup \mathcal{C}))$ (and so $\vdash_{\text{grounded}} = \vdash_{\text{stable}}^{\text{rea}}$).¹

Proof. *Ad 1.* Let $a = \Delta_1^f, \Theta_1, \Gamma_1^c \Rightarrow \Sigma \in \text{Arg}(\mathcal{K})$. **Left-to-Right.** Consider a defeater $b = \Delta_2^f, \Theta_2, \Gamma_2^c \Rightarrow \neg(\varphi, \psi)$ of a . By Proposition 1, $\Theta_2 \cup \{(\varphi, \psi)\}$ is \mathcal{C} -inconsistent in \mathcal{K} . Since $a \in \bigcap \text{stb}(\mathcal{AF})$ and $(\varphi, \psi) \in \text{prem}(a)$, and by Proposition 3, Θ_2 is not contained in a consistent set of norms in \mathcal{K} and it is therefore inconsistent. By Proposition 1, there are $\Delta_3^f \cup \Gamma_3^c \subseteq \mathcal{F} \cup \mathcal{C}$ such that $\Delta_3^f, \Gamma_3^c \Rightarrow \neg\Theta_2$ defeats b . **Right-to-Left.** It is easy to see that $\bigcap \text{stb}(\mathcal{AF})$ contains every argument it defends. Suppose now that $b = \Gamma \Rightarrow \Delta$ is \mathcal{C} -inconsistent. By Proposition 1, there is a $c = \Omega \Rightarrow \neg(\Gamma \cap \mathcal{L}^n)$ that defeats b and for which $\Omega \cap \mathcal{L}^n = \emptyset$. Since c has no defeaters, $c \in \bigcap \text{stb}(\mathcal{AF})$. So, $\bigcap \text{stb}(\mathcal{AF})$ defends a and therefore $a \in \bigcap \text{stb}(\mathcal{AF})$.

Ad 2. **Left-to-Right.** Straightforward. We show **Right-to-Left.** Let $a \in \bigcap \text{stb}(\mathcal{AF})$. By Item 1, a is defended by $\text{Arg}(\mathcal{F} \cup \mathcal{C})$. Clearly, $\text{Arg}(\mathcal{F} \cup \mathcal{C}) \subseteq \mathcal{G}_1 \subseteq \text{grd}(\mathcal{AF})$ since arguments in this set do not have defeaters. So, $a \in \mathcal{G}_2 \subseteq \text{grd}(\mathcal{AF})$. \square

¹Recall that $\bigcup_{i \geq 0} \mathcal{G}_i$ where $\mathcal{G}_0 = \emptyset$ and $\mathcal{G}_{i+1} = \text{Defended}(\mathcal{G}_i)$. The proposition states the computationally interesting result that the fixed-point construction of the grounded extension terminates on the second iteration.

Proposition 4. *Let \mathcal{AF} be a DAC-induced \mathcal{AF} for $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ and let $\mathcal{A} \subseteq \text{Arg}$: \mathcal{A} is a stable extension iff \mathcal{A} is a preferred extension (and so $\vdash_{\text{preferred}}^* = \vdash_{\text{stable}}^*$ for $*$ $\in \{\cup, \cap, \text{rea}\}$).*

Proof. Left-to-Right. Lemma 15 in [13]. **Right-to-Left.** Suppose not, then there is an $a \in \text{Arg}(\mathcal{K}) \setminus \mathcal{A}$ but no $b \in \mathcal{A}$ for which $(b, a) \in \text{Att}$. Since \mathcal{A} is preferred it is conflict-free. By Proposition 2, $\mathcal{N}' = \{(\varphi, \psi) \mid (\varphi, \psi) \in \Delta \cap \mathcal{L}^n \text{ and } \Delta \Rightarrow \Gamma \in \mathcal{A}\}$ is \mathcal{C} -consistent and so $\mathcal{N}' \subseteq \mathcal{N}''$ for some $\mathcal{N}'' \in \text{maxfam}(\mathcal{K})$ and there is a stable extension $\mathcal{A}' = \text{Arg}(\mathcal{F} \cup \mathcal{N}'' \cup \mathcal{C})$. Clearly, $\text{Arg}(\mathcal{F} \cup \mathcal{N}' \cup \mathcal{C}) \subseteq \text{Arg}(\mathcal{F} \cup \mathcal{N}'' \cup \mathcal{C})$ and so, \mathcal{A} is not maximally admissible. Contradiction. \square

4. Dialogues and Contrastive Explanations

We now provide dialogue models for contrastive deontic explanations. A contrastive explanatory dialogue starts with a command “ $\varphi^o!$ ” issued by the explainer, immediately followed by the explainee asking a question of the form: “*Why φ^o , despite ψ^o ?*”

Due to the conflict sensitivity of norm systems [3, 11], we consider contrastive why-questions as the starting point of explanatory episodes [14]. We refer in what follows to φ^o as the *claim* and to ψ^o as the *counter-claim*.² We do not assume that φ^o and ψ^o are derivable from the given normative system \mathcal{K} , nor do we assume that there is a dialectical relation between φ^o and ψ^o , referred to as the *contrastive link*. Both must become explicit (if existent) through the dialogue itself. We say there exists a contrastive link when two arguments a and b exist concluding φ^o , respectively ψ^o , which are incompatible, meaning that there is no stable extension containing both. In such a case, an incompatibility argument can be provided using the premises in a and b :

Proposition 5. *Let $\mathcal{K} = \langle \mathcal{F}, \mathcal{N}, \mathcal{C} \rangle$ and $\mathcal{AF}(\mathcal{K}) = \langle \text{Arg}, \text{Att} \rangle$. For any two arguments $a = \Delta \Rightarrow \varphi^o$, $b = \Gamma \Rightarrow \psi^o \in \text{Arg}$: there is no stable extension \mathcal{S} with $a, b \in \mathcal{S}$ iff there is a DAC-derivable argument $\Delta, \Gamma, \Omega \Rightarrow$ with $\Omega \subseteq \mathcal{C}$ (we call a and b \mathcal{C} -incompatible).*

Proof. Left-to-Right. Let $\mathcal{N}' = (\Gamma \cup \Delta) \cap \mathcal{L}^n$, and $\mathcal{F}' = \{\varphi \mid \varphi^f \in \Gamma \cup \Delta\}$. By Prop. 3, there is no $\mathcal{M} \in \text{maxfam}(\mathcal{K})$ with $\mathcal{N}' \subseteq \mathcal{M}$. So, there is a $\Omega \subseteq \mathcal{C}$ for which $\perp \in \text{Cn}(\text{out}_3(\mathcal{N}', \mathcal{F}') \cup \Omega)$. By Prop. 1 and a **Cut** application, $\vdash_{\text{DAC}} \mathcal{N}', \mathcal{F}', \Omega \Rightarrow$. **Right-to-Left.** Straightforward. \square

An explanatory dialogue addressing “*Why φ^o , despite ψ^o ?*” is successful whenever it contains

- c1** an argument a for φ^o and a demonstration that all (indirect) objections to a can be met;
- c2** an argument b for ψ^o such that a and b are \mathcal{C} -incompatible (recall Prop. 5);
- c3** an argument c defeating b and a demonstration that all (indirect) objections to c can be met;
- c4** a demonstration that the demonstrations in **c1** and **c3** are \mathcal{C} -compatible.

Informally, **c1** provides the ‘illative explanation’ of “ $\varphi^o!$ ” by stating a containing the facts and norms in view of which φ^o holds. It also provides the ‘dialectic explanation’ of φ^o by refuting all

²In the philosophical literature deontic explanations are relatively unexplored. Our account takes the question-oriented pragmatic approach to contrastive explanation (cf. [14]), which naturally extends to dialogue models. It accords with [15] who calls upon defeasible moral principles (here interpreted as norms) to substantiate explanations and with [16] who takes defeasible norms to serve as justifications, namely, norms ground as to why the called-upon facts are explanatory. See also [17] for the role of justification in the context of normative explanations.

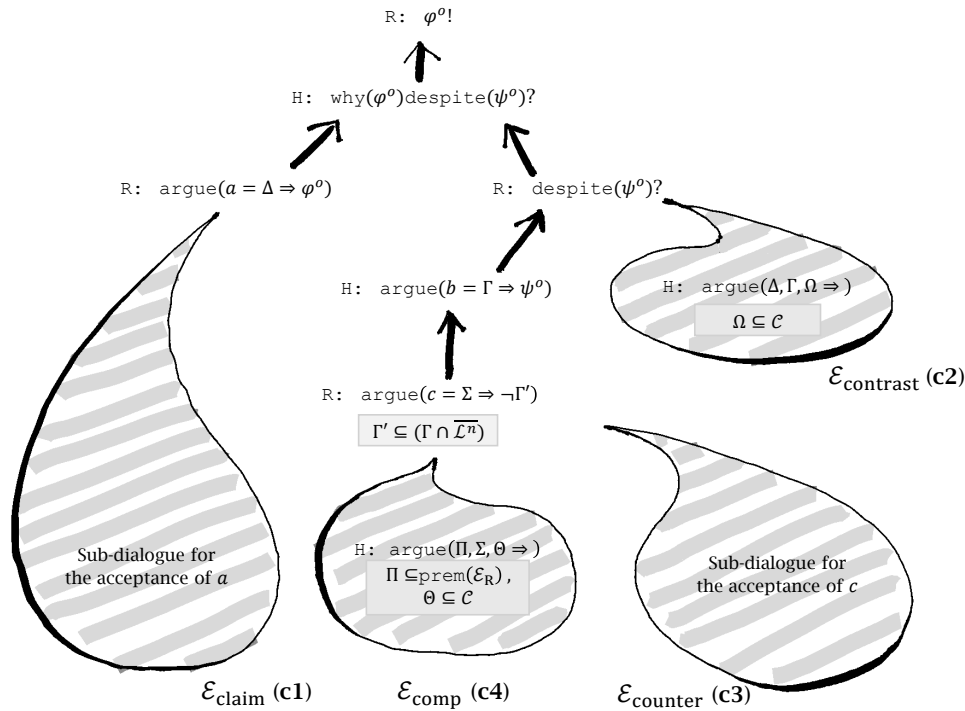


Figure 3: A DAC-based contrastive explanatory dialogue (CED) for explainer (R) and explainee (H).

possible objections the explainee may have against concluding φ^o . **c2** makes explicit the contrast between the argument for φ^o with an argument for ψ^o , and **c3** provides illative and dialectical explanations for why ψ^o can be successfully objected to (our terminology mirrors Johnson’s well-known two-tier model of argument [18]). Last, **c4** ensures that the two sub-explanations in **c1** and **c3** form a \mathcal{C} -compatible view. The intuitive idea of contrastive explanatory dialogues, following **c1-c4**, is provided in Figure 3. Below, we make this formally precise.

Although explanatory dialogues are collaborative, we assume a *burden of proof* for the explainer with respect to **c1** and **c3**, and for explainee with respect to **c2** and **c4**. For the sake of simplified reference, we call the explainee ‘human’ (H) and the explainer ‘robot’ (R).

Explanatory Dialogues Let $\mathcal{AF}(\mathcal{K})$ be a DAC-induced argumentation framework. Let R be the explainer and H the explainee. A *contrastive explanatory dialogue* (CED) is a sequence $\mathcal{E} = \langle m_1, \dots, m_n \rangle$ of tuples $m_i = \langle \text{pl}, \text{lo}, \text{ta} \rangle$ called *moves* such that i is m_i ’s position in the dialogue, $\text{pl}(m_i) \in \{R, H\}$ is the player making move m_i , $\text{lo}(m_i) \in \text{Locutions}$ is the locution in m_i , and $\text{ta}(m_i) \in \{m_1, \dots, m_{i-1}\} \cup \{\emptyset\}$ is the target of m_i . $\text{Locutions} = \{(x)!, \text{why}(x)\text{despite}(y)?, \text{despite}(x)?, \text{argue}(z)\}$ is the set of expressions that interlocutors may use, where x and y range over \mathcal{L}^o and z ranges over $\text{Arg}(\mathcal{K})$. \mathcal{E} is a sem-CED (for $\text{sem} \in \{\text{preferred}, \text{grounded}\}$) whenever \mathcal{E} satisfies the *protocol* stipulated by **P1-P5**.

P1 Dialogue Commencement Rules for $m_i \in \mathcal{E}$ with $i \in \{1, 2, 3, 4\}$:

$$\begin{aligned} m_1 &= \langle R, \varphi^o!, \emptyset \rangle & m_3 &= \langle R, \text{argue}(a), m_2 \rangle \text{ with } \text{conc}(a) = \varphi^o \\ m_2 &= \langle H, \text{why}(\varphi^o)\text{despite}(\psi^o)?, m_1 \rangle & m_4 &= \langle R, \text{despite}(\psi^o)?, m_2 \rangle \end{aligned}$$

P2 General Rules for each $m_i, m_k \in \mathcal{E}$:

- i) if $i > 4$, then $\text{lo}(m_i) = \text{argue}(x)$, and $\text{ta}(m_i) = m_k$ with $k < i$.
- ii) if $\text{ta}(m_i) = m_k$, $\text{lo}(m_i) = \text{argue}(x)$ and $\text{lo}(m_k) = \text{argue}(y)$, then $\text{conc}(x) = \neg\Delta \in \overline{\mathcal{L}^n}$ for some $\Delta \subseteq \text{prem}(y)$;
- iii) if $\text{ta}(m_i) = \text{ta}(m_k)$ and $i \neq k$, then $\text{lo}(m_i) \neq \text{lo}(m_k)$;
- iv) if $\text{ta}(m_i) = m_k$, then $\text{pl}(m_i) \neq \text{pl}(m_k)$.

P1 stipulates that (1) R starts the dialogue with a command to which, (2) H responds with a contrastive why-question. Then, (3) R must provide reasons for the command and, after that, (4) shifts the burden of proof to H requesting support for the contrastive claim. **P2** stipulates rules that hold for both R and H: (i) after the start of the dialogue any player may continue making moves that target previous moves by stating arguments, where (ii) arguments moved against other arguments express undermining defeats, (iii) players may not move an argument twice against the same move, and (iv) they may not attack their own claims.

P3 Explainer Rules for each $m_i, m_k \in \mathcal{E}$, if $\text{ta}(m_i) = m_2$, then $i = 3$ or $i = 4$.

P4 Explainee Rules for each $m_i, m_j, m_k \in \mathcal{E}$:

- i) if $\text{ta}(m_i) = \text{ta}(m_j) = \text{ta}(m_k) = m_4$ (and, so, $\text{pl}(m_i) = \text{H}$), then $|\{i, j, k\}| \leq 2$.
- ii) if $\text{ta}(m_i) = \text{ta}(m_j) = m_4$, $i \neq j$, $\{\text{lo}(m_i), \text{lo}(m_j)\} = \{\text{argue}(b), \text{argue}(c)\}$, $\text{lo}(m_3) = \text{argue}(a)$, then $\text{conc}(b) = \psi^o$ and $c = \text{prem}(a)$, $\text{prem}(b)$, $\Omega \Rightarrow$ with $\Omega \subseteq \mathcal{C}$.
- iii) if $\text{ta}(m_i) = m_k$, $\text{ta}(m_k) = m_j$, $\text{ta}(m_j) = m_4$, $\text{lo}(m_j) = \text{argue}(b)$, with $\text{conc}(b) \neq \emptyset$, and $\text{lo}(m_k) = \text{argue}(d)$, then
 - either $\text{lo}(m_i) = \text{argue}(\Sigma, \text{prem}(d), \mathcal{C} \Rightarrow)$, and $\Sigma \subseteq \text{prem}(\mathcal{E}_R)$ with $\mathcal{E}_R = \{a \mid m \in \mathcal{E}, \text{pl}(m) = \text{R}, \text{lo}(m) = \text{argue}(a)\}$;
 - or $\text{lo}(m_i) = \text{argue}(e)$ for some e with $\text{conc}(e) = \neg\Delta \in \overline{\mathcal{L}^n}$ and $\Delta \subseteq \text{prem}(d)$.

P3 states that R must provide *exactly* two moves against the contrastive why-question (one of which provides reasons, and the other questioning the contrastive claim). **P4** stipulates that the explainee H may (i) make *at most* two moves against R's questioning of the contrastive link, (ii) one of which is an argument providing reasons for the counter-claim, one which shows the \mathcal{C} -incompatibility of the arguments for the claim and the counter-claim ψ^o . Then, (iii) H may also move against the explainer's argument d opposing the reasons for the counter-claim. In case d is incompatible with the other arguments offered by R, R engages in incoherent reasoning. H may, thus, oppose by demonstrating the \mathcal{C} -incompatibility of d and other R arguments.

A CED has a tree-structure since each move has exactly one predecessor (except for the root). A *branch* of \mathcal{E} containing m_i is the *maximal linear sequence* $\text{branch}(m_i) = \langle m_{j_1}, \dots, m_{i=j_k}, \dots, m_{j_n} \rangle$ such that for each m_{j_l} and $m_{j_{l+1}}$, $\text{ta}(m_{j_{l+1}}) = j_l$. We say m_{j_n} is a leaf. Each CED consists of four subdialogues (see Fig. 3), which constitute four sub-explanations: a subdialogue $\mathcal{E}_{\text{claim}}$ (cf. **c1**) that engages with the argument a given in favour of the claim φ^o (generated from m_3 down); a subdialogue $\mathcal{E}_{\text{counter}}$ (cf. **c3**) that engages with the argument b given in favour of the counter-claim ψ^o (generated from the move attacking the argument providing reasons for ψ^o); and the subdialogues $\mathcal{E}_{\text{contrast}}$ (cf. **c2**) and $\mathcal{E}_{\text{comp}}$ (cf. **c4**) each containing at most one node with an argument that shows the \mathcal{C} -incompatibility of a and b , respectively, joint

\mathcal{C} -incompatibility of R's explanations $\mathcal{E}_{\text{claim}}$ and $\mathcal{E}_{\text{counter}}$. These four subdialogues determine when a given dialogues is (un/semi)successful.

Before defining success, we add **P5** to the protocol to accommodate reasoning with preferred and grounded acceptance of arguments. For $\text{sem} \in \{\text{preferred}, \text{grounded}\}$, (i) R may move at most one counter-argument to each H argument. For preferred dialogues, (ii) H is not allowed to move the same argument twice on a branch in $\mathcal{E}_{\text{claim}}$ or $\mathcal{E}_{\text{counter}}$. For grounded dialogues, (iii) R is not allowed to move the same argument twice on a branch. We note that (i)-(iii) follow the protocols for admissible (and, so, preferred) and grounded argumentation games [10].

P5 Preferred and Grounded Rules for each $m_i, m_j \in \mathcal{E}$, and $\text{sem} \in \{\text{preferred}, \text{grounded}\}$:

- i) $\text{ta}(m_i) = \text{ta}(m_k) = m_j$ with $j > 3$ and $\text{pl}(m_i) = \text{R}$, then $m_i = m_k$;
- ii) if $\text{sem} = \text{preferred}$, $\text{pl}(m_i) = \text{H}$, and $\text{ta}(m_i) = m_j$, then there is no $m_k \in \text{branch}(m_j)$ for which $\text{pl}(m_k) = \text{H}$, $\text{lo}(m_i) = \text{lo}(m_k)$ and $i \neq k$;
- iii) if $\text{sem} = \text{grounded}$, $\text{pl}(m_i) = \text{R}$, and $\text{ta}(m_i) = m_j$, then there is no $m_k \in \text{branch}(m_j)$ for which $\text{pl}(m_k) = \text{R}$, $\text{lo}(m_i) = \text{lo}(m_k)$ and $i \neq k$.

Successful dialogues Let $\mathcal{AF}(\mathcal{K})$ be DAC-induced. A CED \mathcal{E} satisfying **P1-P5** is:

- *successful* if $\mathcal{E}_{\text{contrast}} \neq \emptyset$, $\mathcal{E}_{\text{comp}} = \emptyset$, and $\mathcal{E}_{\text{claim}}$ and $\mathcal{E}_{\text{counter}}$ both contain R-leaves only;
- *semi-successful* if $\mathcal{E}_{\text{contrast}} = \emptyset = \mathcal{E}_{\text{comp}}$, and $\mathcal{E}_{\text{claim}}$ contains R-leaves only;
- *unsuccessful* if neither of the above holds.

Then, \mathcal{E} is sem-successful when it is *saturated* (i.e., all movable arguments from $\mathcal{AF}(\mathcal{K})$ are moved in \mathcal{E} ; cf. [9]) and \mathcal{E} is successful (similar for semi- and unsuccessful).

In brief, a successful CED features **(c1)** an illative explanation that is supplemented by a dialectical explanation ($\mathcal{E}_{\text{claim}}$ contains only R-leaves), where **(c2)** the explainee is able to demonstrate the incompatibility of the contrastive claim ($\mathcal{E}_{\text{contrast}} \neq \emptyset$), the latter which **(c3)** the explainer successfully counters ($\mathcal{E}_{\text{counter}}$ contains only R-leaves). Furthermore, **(c4)** the position taken by R in $\mathcal{E}_{\text{claim}}$ and $\mathcal{E}_{\text{counter}}$ must be \mathcal{C} -compatible. A semi-successful dialogue features **(c1)**, but H is not able to demonstrate the adequacy of the contrastive link ($\mathcal{E}_{\text{contrast}} = \emptyset$). A dialogue can be unsuccessful for various reasons, e.g., R cannot provide an illative or dialectic explanation, or R cannot argue against H's counter-claim.

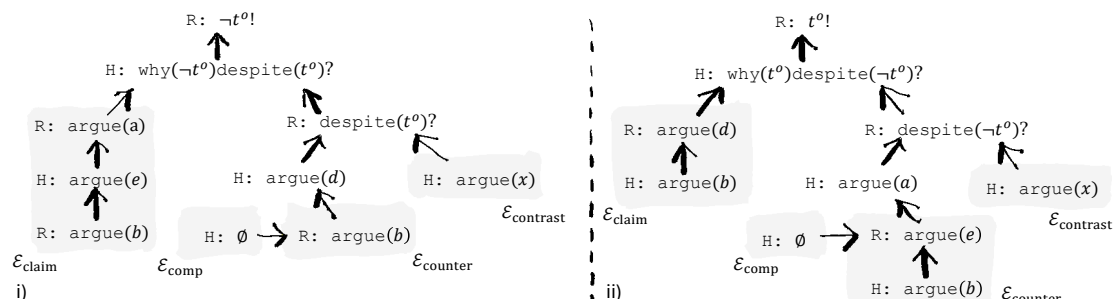
Under saturation, it can be easily checked that the sub-dialogues in $\mathcal{E}_{\text{claim}}$ and $\mathcal{E}_{\text{counter}}$ are for $\text{sem} \in \{\text{preferred}, \text{grounded}\}$ instances of credulous preferred and grounded argumentation games [10] (where for the latter credulous equals skeptical entailment). Hence, we obtain dialogue models that construct explanations for credulous I/O entailment (i.e., when $\text{sem} = \text{preferred}$) and for skeptical I/O entailment under shared reasons (i.e., when $\text{sem} = \text{grounded}$).

Proposition 6. Let $\mathcal{AF}(\mathcal{K}) = \langle \text{Arg}, \text{Att} \rangle$ be DAC-induced, $\text{sem} \in \{\text{preferred}, \text{grounded}\}$, and $\mathcal{E}_* = \langle m_i, \dots, m_n \rangle$ be $\mathcal{AF}(\mathcal{K})$ -based with $\text{lo}(m_i) = \text{argue}(a)$ and $*$ $\in \{\text{claim}, \text{counter}\}$:

- if \mathcal{E}_* is saturated and contains only R leaves then $a \in \mathcal{S}$ for some sem-extension;
- if $a \in \mathcal{S}$ for some sem-extension \mathcal{S} , there is a saturated extension of \mathcal{E}_* with only R leaves.

Proof. Straightforward modification of the proofs of Theorems 6.2 and 6.5 in [10]. \square

Example 3. Let $\text{sem} \in \{\text{preferred}, \text{grounded}\}$: Figure i) provides a successful saturated sem -CED for “Why $\neg t^o$, despite t^o ?” notice that $\mathcal{E}_{\text{comp}}$ is empty since $\mathcal{E}_R = \{a, b\}$ is \mathcal{C} -compatible. Figure ii) contains an unsuccessful saturated sem -CED for “Why t^o , despite $\neg t^o$?” where H refutes the claim (with b) and defends the counter-claim (with a and b). The arguments in i) and ii) reference those in Fig. 2 of Ex. 2 (for space reasons, we adopted an example for which grounded equals stable).



An alternative successful CED for i) exists that includes R moving g against e , followed by H moving f , which is then defeated by R moving b . However, here we can use Proposition 3-1, giving us for $\text{sem} = \text{grounded}$ the existence of a strategic shortcut by directly moving argument b against e .

5. Challenges for Dialogical Deontic Explanations

This paper shows how to incorporate existing results in formal argumentation and refine them to yield contrastive explanatory dialogues (CEDs) in the context of defeasible normative reasoning, Input/Output logic in particular. This work is the first of its kind and, so, we end by highlighting some key challenges for deontic explanations (through formal argumentation).

Challenge 1: Conflict Types The contrastive claims offered by the explainee may give rise to various kinds of conflicts with the main claim. Two particularly interesting cases when dealing with (conditional) norms are specificity (you are not allowed to park, unless you are medical personnel) and contrary-to-duty (don’t be late, but if you are, not more than 10 minutes). Good explanations should make transparent the type of conflicts involved.

Challenge 2: Cognitive Adequacy A good explainer seeks to understand the explainee in order to tailor the given explanation to precisely target the gaps in the explainee’s understanding. For this the explainer may use queries and strategic argumentation, complemented by a theory of (the explainee’s) mind. Moreover, the knowledge bases of the explainer and the explainee may be disjoint and incomplete. Tailored explanations must additionally keep track of commitments and shifts therein throughout a dialogue.

Challenge 3: Richer Handling of Contrastives The explainee may offer contrastive claims that are, under thorough analysis, not really incompatible with the offered claim. A good explainer should catch such cases and provide an argument concerning the compatibility of the claims. For this, more proof-theoretic resources have to be developed. In such cases, the explainee should be able to withdraw or replace the contrast.

Challenge 4: Richer Deontic Vocabulary Often, normative codes are richer than the ones studied here, e.g., they may contain priority orderings over norms and permissive norms.

These come with challenges, for instance concerning reinstatement (e.g., permissions generally do not reinstate obligations). Dialogues ideally accommodate such complexity.

Challenge 5: Casuistry In many application contexts of ethical (e.g., in bioethics) and legal reasoning, we find case-based reasoning when reasoning towards obligations, rights, and permissions. Deontic explanations of such conclusions need a different conceptual base than the one provided here, posing their own specific challenges (e.g., balancing reasons).

Acknowledgements. This work was partially funded by the “Logical Methods of Deontic Explanations” (LoDeX) project, Deutsche Forschungsgemeinschaft, Project number 511915728.

References

- [1] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38.
- [2] O. Arieli, A. Borg, J. Heyninck, C. Straßer, Logic-based approaches to formal argumentation, in: D. Gabbay, M. Giacomin, G. R. Simari, M. Thimm (Eds.), *Handbook of Formal Argumentation, Volume 2*, College Publications, 2021, pp. 1793–1898.
- [3] D. Makinson, L. van der Torre, Constraints for Input/Output logics, *Journal of Philosophical Logic* 30 (2001) 155–185.
- [4] J. F. Horty, *Reasons as defaults*, Oxford University Press, 2012.
- [5] K. van Berkel, C. Straßer, Reasoning with and about norms in logical argumentation, in: *Proceedings of COMMA 2022*, volume 353, IOS press, 2022, pp. 332 – 343.
- [6] K. van Berkel, C. Straßer, Z. Zhou, Towards an argumentative unification of default reasoning, in: *Proceeding of COMMA 2024*, IOS press, 2024, p. TBA.
- [7] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, *Argumentative XAI: A survey*, 2021.
- [8] L. Amgoud, N. Maudet, S. Parsons, Modelling dialogues using argumentation, in: *Proceedings Fourth International Conference on MultiAgent Systems*, IEEE, 2000, pp. 31–38.
- [9] H. Prakken, Coherence and flexibility in dialogue games for argumentation, *Journal of Logic and Computation* 15 (2005) 1009–1040.
- [10] S. Modgil, M. Caminada, Proof theories and algorithms for abstract argumentation frameworks, in: *Argumentation in artificial intelligence*, Springer, 2009, pp. 105–129.
- [11] D. Gabbay, J. F. Horty, X. Parent, R. van der Meyden, L. van der Torre, *Handbook of Deontic Logic and Normative Systems, Volume 1*, College Publications, United Kingdom, 2013.
- [12] O. Arieli, K. van Berkel, C. Straßer, Defeasible normative reasoning: A proof-theoretic integration of logical argumentation, in: *Proceedings of AAI 2024*, 2024, pp. 10450–10458.
- [13] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Art. Int.* 77 (1995) 321–357.
- [14] P. Lipton, Contrastive explanation, *Royal Inst. of Philosophy Suppl.* 27 (1990) 247–266.
- [15] U. D. Leibowitz, Scientific explanation and moral explanation, *Noûs* 45 (2011) 472–503.
- [16] M. Scriven, Explanations, predictions, and laws, in: *Minnesota Studies in the Philosophy of Science, Vol. 3*, University of Minnesota Press, Minneapolis, 1962, pp. 170–230.
- [17] P. Väyrynen, Normative explanation and justification, *Noûs* 55 (2021) 3–22.
- [18] R. H. Johnson, *Manifest rationality: A pragmatic theory of argument*, Routledge, 2000.