Stable Normative Explanations: From Argumentation to Deontic Logic

Stable Normative Explanations: From Argumentation to Deontic Logic

Cecilia Di Florio[1], Guido Governatori[2][0000−0002−9878−2762], Antonino Rotolo[1][0000−0001−5265−0660], and Giovanni Sartor[3][0000−0001−5680−0080]

[1] ALMA AI and Department of Legal Studies, University of Bologna, Italy {cecilia.diflorio2,antonino.rotolo}@unibo.it
[2] Cooroibah, QLD 4565, Australia guido@governatori.net
[3] ALMA AI/Department of Legal Studies, University of Bologna, and EUI, Italy giovanni.sartor@unibo.it

Abstract. This paper examines how a notion of stable explanation developed elsewhere in Defeasible Logic can be expressed in the context of formal argumentation. With this done, we discuss the deontic meaning of this reconstruction and show how to build from argumentation neighborhood structures for deontic logic where this notion of explanation can be characterised. Some direct complexity results are offered.

1 Introduction

Resorting to machine learning to predict outcomes in legal proceedings is very much discussed in the literature as well as by policy-makers (for an overview, see, e.g., [21,6,4]). Indeed, such techniques can be used for algorithmic decision predictors to support judges in individual cases, to assist litigants in estimating their likelihood of winning a case or in examining various biases on legal decision-making processes [6]. One of the most challenging contexts in which to introduce AI is within courts. Judges are often reluctant to adopt these tools for two reasons: (a) it could undermine the independent exercise of judicial power, and (b) AI is anything but transparent and explainable.

Developing Explainable-AI systems is thus more and more important in the law since '*transparency*' and '*justification*' of legal decision-making both require formalising normative explanations [1]. Normative explanation is a type of explanation where norms (in addition to factual information) are crucial: if reframed in the context of legal decision-making, this means to explain why a legal conclusion (such as an obligation) ought to be the case on the basis of certain norms (such as one prescribing to compensate for the damages for which we are liable) and facts (such as the fact that I causally contributed to cause a damage) [2,24]. In the context of judicial reasoning, the idea of normative explanation is now emerging in the literature (see [12,26,19,18]).

Legal proceedings are adversarial in nature. In this perspective, if a judge or a litigant aim at predicting possible outcomes, this fact must be taken into account, and formal tools to make such predictions understandable should allow for checking if a certain legal outcome is *stable* [11,12,22]. This is especially true in an argumentation perspective, where the adversarial structure of proceedings become more transparent. In

such a perspective, given some facts, the proceeding aims at determining what legal requirements hold, and whether such legal requirements have been fulfilled. (In)Stability means that, if more/new facts were presented, the outcome of a case might be quite different or can even be modified. How to ensure a specific outcome for a case, which, in an adversarial setting, can be understood as addressing the question of how to ensure that the facts presented by a party are 'resilient' to the attacks from the opponent?

In this paper we adopt [11,12]'s definition of stability and elaborate it in the argumentation setting of Defeasible Logic [3]. Apart from some details, while valuable, this extension is technically rather straightforward. However, we are interested in second, and more challenging, research question: *What is the deontic meaning of stable normative explanation as developed in an argumentation setting*? In fact, in legal argumentation, a typical outcome of judicial decisions are obligations and permissions.

In moving to the deontic domain, we must notice that deontic argumentation can be developed in various ways [16,28]. As commonly done in the AI&Law literature [27], we assume that legal norms are rules having the form $\phi_1, \ldots, \phi_n \Rightarrow \psi$ and we follow this intuition:

Intuition 1 *Let* AF *be an argumentation framework where arguments are built using rules of the form* $\phi_1, \ldots, \phi_n \Rightarrow \psi$. *Then* $\mathbf{OBL}\psi$ *holds in* AF *iff* ψ *is justified w.r.t.* AF.

Once we have defined the argumentative setting and identified some notions of normative explanation, we adapt [15]'s method and show how *this machinery can be reconstructed in neighborhood semantics for classical deontic logics [8] and how the notion of explanation can be semantically characterised*.

The layout of article is as follows. Section 2 recalls the basics of Defeasible Logic and offers a variant of the idea of argumentation framework based on such a logic. Section 3 presents the definitions of normative explanation and stable normative explanation. Section 4 illustrates how to move from argumentation structures to neighbourhood semantics for deontic logic. Section 5 applies the ideas of Sections 3 and 4 to semantically reconstruct the concept of normative explanation.

2 Background

2.1 Defeasible Logic

The logical apparatus we utilise is the standard Defeasible Logic (DL) [3]. In this section we present the basics of DL.

Let PROP be the set of propositional atoms, then the set of literals $\text{Lit} = \text{PROP} \cup \{\neg p \mid p \in \text{PROP}\}$. The *complementary* of a literal p is denoted by $\sim p$: if p is a positive literal q then $\sim p$ is $\neg q$, if p is a negative literal $\neg q$ then $\sim p$ is q. Literals are denoted by lower-case Roman letters. Let Lab be a set of labels to represent names of rules.

A *defeasible theory* D is a tuple $(F, R, >)$, where F is the set of facts (indisputable statements), R is the rule set, and $>$ is a binary relation over R.

R is partitioned into three distinct sets of rules, with different meanings to draw different "types of conclusions". *Strict rules* are rules in the classical sense: whenever the

premises are the case, so is the conclusion. We then have *defeasible rules* which represent the non-monotonic part (along which defeaters) of the logic: if the premises are the case, then typically the conclusion holds as well unless we have contrary evidence that opposes and prevents us from drawing such a conclusion. Lastly, we have *defeaters*, which are special rules whose purpose is to prevent contrary evidence from being the case. It follows that in DL, through defeasible rules and defeaters, we can represent in a natural way exceptions (and exceptions to exceptions, and so forth).

We finally have the superiority relation $>$, a binary relation among couples of rules that is the mechanism to solve conflicts. Given the two rules r and t, we have $\langle r,t\rangle \in >$ (or simply $r > t$), in the scenario where both rules may fire (can be activated), r's conclusion will be preferred to t's.

In general, a rule $r \in R$ has the form $r\colon A(r) \hookrightarrow C(r)$, where: (i) $r \in$ Lab is the unique name of the rule, (ii) $A(r) \subseteq$ Lit is r's (set of) antecedents, (iii) $C(r) = l \in$ Lit is its conclusion, and (iv) $\hookrightarrow \in \{\rightarrow, \Rightarrow, \rightsquigarrow\}$ defines the type of rule, where: $\rightarrow$ is for strict rules, $\Rightarrow$ is for defeasible rules, and $\rightsquigarrow$ is for defeaters.

Some standard abbreviations. R_s denotes the set of strict rules in R, and the set of strict and defeasible rules is denoted by Rs; $R[l]$ denotes the set of all rules whose conclusion is l.

A *conclusion* of D is a *tagged literal* with one of the following forms:

$+\Delta l$ (resp. $-\Delta l$) means that l is *definitely proved* (resp. *strictly refuted/non provable*) in D, i.e., there is a definite proof for l in D (resp. a definite proof does not exist).
$+\partial l$ (resp. $-\partial l$) means that l is *defeasibly proved* (resp. *defeasibly refuted*) in D, i.e., there is a defeasible proof for l in D (resp. a definite proof does not exist).

The definition of proof is also the standard in DL. Given a defeasible theory D, a proof P of length n in D is a finite sequence $P(1), P(2), \ldots, P(n)$ of tagged formulas of the type $+\Delta l$, $-\Delta l$, $+\partial l$, $-\partial l$, where the proof conditions defined in the rest of this section hold. $P(1..n)$ denotes the first n steps of P.

All proof tags for literals are standard in DL [3]. We present only the positive ones as the negative proof tags can be straightforwardly obtained by applying the *strong negation principle* to the positive counterparts. The strong negation principle applies the function that simplifies a formula by moving all negations to an innermost position in the resulting formula, replaces the positive tags with the respective negative tags, and the other way around, see [13].

Positive proof tags ensure that there are effective decidable procedures to build proofs; the strong negation principle guarantees that the negative conditions provide a constructive method to verify that a derivation of the given conclusion is not possible.

The definitions of $\pm\Delta$ describe forward-chaining of strict rules and are omitted.

Defeasible derivations are based on the notions of a rule being applicable or discarded. A rule is *applicable* at a given derivation step when every antecedent has been proved at any previous derivation step. Symmetrically, a rule is *discarded* when at least one antecedent has been previously refuted.

Definition 1 (Applicable & Discarded).
Given a defeasible theory D, a literal l, and a proof $P(n)$, we say that

- $r \in R[l]$ *is* applicable *at* $P(n+1)$ *iff* $\forall a \in A(r).\ +\partial a \in P(1..n)$.
- $r \in R[l]$ *is* discarded *at* $P(n+1)$ *iff* $\exists a \in A(r).\ -\partial a \in P(1..n)$.

Note that a strict rule can be used to derive defeasible conclusions when it is applicable and at least one of its premises is defeasibly but not strictly proved.

Definition 2 $(+\partial)$**.**

$+\partial l$: *If* $P(n+1) = +\partial l$ *then either*
(1) $+\Delta l \in P(1..n)$*, or*
(2.1) $-\Delta{\sim}l \in P(1..n)$*, and*
(2.2) $\exists r \in R[l]$ *applicable s.t.*
(2.3) $\forall s \in R[{\sim}l]$ *either*
(2.3.1) s discarded, or
(2.3.2) $\exists t \in R[l]$ *applicable s.t.* $t > s$.

A literal is defeasibly proved if (1) it has already proved as a strict conclusion, or (2.1) the opposite is not and (2.2) there exists an applicable, defeasible or strict, rule such that any counter-attack is either (2.3.1) discarded or (2.3.2) defeated by an applicable, stronger rule supporting l. Note that, whereas s and t may be defeaters, *r may not*, as we need a strict or defeasible, applicable rule to draw a conclusion.

The last notions introduced in this section are those of extension of a defeasible theory. Informally, an extension is everything that is derived and disproved.

Definition 3 (Theory Extension). *Given a defeasible theory D, we define the set of positive and negative conclusions of D as its* extension*:* $E(D) = (+\Delta, -\Delta, +\partial, -\partial)$*, where* $\pm\# = \{l \mid l \text{ appears in } D \text{ and } D \vdash \pm\#l\}$, $\# \in \{\Delta, \partial\}$.

Theorem 1. *[20] Given a defeasible theory D, its extension* $E(D)$ *can be computed in time polynomial to the size of the theory.*

2.2 Argumentation in Defeasible Logic

Argumentation frameworks for DL and the corresponding argumentation semantics have been in general studied in [10]. Here, we present a variant of it, which is based on a fragment of DL without strict rules and defeaters. Also, since rules are meant to express norms, facts have a special status here, i.e.—as argued in [11,12]—they are meant to capture purely factual information and thus do not occur in the heads of rules (which are supposed to lead to normative conclusions).

Definition 4. *An* argumentation theory *D is a defeasible theory* $(F, R, >)$ *where*

- *R is a (finite) set of defeasible rules,*
- $F \subseteq$ Lit *is a finite consistent set of facts where, for each* $p \in F$, $R[p] \cup R[{\sim}p] = \emptyset$*, and*
- $> \subseteq R \times R$ *is a superiority relation on R.*

By combining the rules in a theory, we can build arguments (we adjust the definition in [25] to meet Definition 4). In what follows, for a given argument A, Conc returns its conclusion, Sub returns all its sub-arguments, Rules returns all the rules in the argument and, finally, TopRule returns the last inference rule in the argument.

Definition 5 (Argument). *Let $D = (F, R, >)$ be an argumentation theory. An argument A constructed from D has either the the form $\Rightarrow_F \phi$ (*factual argument*), where $\phi \in F$, or the form $A_1, \ldots, A_n \Rightarrow_r \phi$ (*plain argument*), where $1 \leq k \leq n$, and*

- *A_k is an argument constructed from D, and*
- *$r : \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \phi$ is a rule in R.*

With regard to argument A, the following holds:

$$\begin{aligned}
&\mathsf{Conc}(A) = \phi \\
&\mathsf{Sub}(A) = \mathsf{Sub}(A_1), \ldots \mathsf{Sub}(A_n), A \\
&\mathsf{TopRule}(A) = r : \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow_r \phi \\
&\mathsf{Rules}(A) = \mathsf{Rules}(A_1), \ldots, \mathsf{Rules}(A_n), \mathsf{TopRule}(A).
\end{aligned}$$

We say that any arguments $\Rightarrow_F \phi$ or $A_1, \ldots, A_n \Rightarrow_r \phi$ are arguments for ϕ.

The following standard definitions are from [10].

Definition 6 (Attack, support, and undercut). *A plain argument A* attacks *a plain argument B if a conclusion of A is the complement of a conclusion of B. We define the* attack relation $\gg$ *such that, for any arguments A and B, $\langle A, B\rangle \in \gg$ (or, in short, $A \gg B$) iff A attacks B. A set of plain arguments* arg *attacks a plain argument B if there is an argument A in* arg *that attacks B.*

A proper subargument $B = \mathsf{Sub}(A)$ of an argument A is such that $B \neq A$.

An argument A is supported *by a set of arguments* arg *if every proper subargument of A is in* arg.

An argument A is undercut *by a set of arguments* arg *if* arg *supports an argument B attacking a proper subargument of A.*

Notice that conflicts between arguments only consider plain arguments: arguments of the form $\Rightarrow_F \phi$ can be ignored because the set of facts is assumed to be consistent and no fact (or its negation) can occur in the head of any rule [11]. The definition above from [10] does not make any reference to the superiority relation, since it is easy to see that the current semantics is a special case of that of [9] when the superiority relation is empty (and for every argumentation theory can be transformed into an equivalent one without the superiority relation is empty). However, the superiority relation can be taken into account by incorporating it into the definition of attack. In other words, for any argument A, if $\mathsf{TopRule}(A) = r : \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow_r \phi$, A attacks another argument B if, and only if $\mathsf{TopRule}(A)$ is stronger than $\mathsf{TopRule}(B)$.

We can now define the argumentation framework.

Definition 7 (Argumentation Framework). *Let $D = (F, R, >)$ be an argumentation theory. The argumentation framework $\mathsf{AF}(D)$ determined by D is $(\mathscr{A}, \gg)$ where $\mathscr{A}$ is the set of all arguments constructed from D, and $\gg$ is the attack relation defined above.*

Definition 8 (Acceptable and rejected argument). *Let $D = (F, R, >)$ be an argumentation theory and $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D. An argument A in $\mathsf{AF}(D)$ for ϕ is* acceptable *w.r.t to a set of argument* arg *in $\mathsf{AF}(D)$ if A is finite and every argument attacking A is undercut by* arg.

An argument A is rejected *by sets of arguments* arg *and* arg$'$ *in $\mathsf{AF}(D)$ when a proper subargument B of A is in* arg *or B is attacked by an argument supported by* arg$'$.

Definition 9 (Sets of acceptable and rejected arguments). *Let $D = (F, R, >)$ be an argumentation theory and $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D. We define J_i^D as follows.*

- $J_0^D = \emptyset$
- $J_{i+1}^D = \{A \in \mathscr{A} \mid A \text{ is acceptable w.r.t. } J_i^D\}$

The set of justified arguments *in an argumentation theory D is $JArgs^D = \cup_{i=1}^{\infty} J_i^D$.*

We define R_i^D as follows.

- $R_0^D = \emptyset$
- $R_{i+1}^D = \{B \in \mathscr{A} \mid B \text{ is rejected by } R_i^D \text{ and } JArgs^D\}$

The set of rejected arguments *in an argumentation theory D is $RArgs^D = \cup_{i=1}^{\infty} R_i^D$.*

$JArgs^D$ corresponds the *extension of the argumentation framework* determined by D.

The following are thus standard result that can be obtained:

Theorem 2. *Let D be an argumentation theory. Then,*

- *an argument A and its conclusion ϕ are justified w.r.t. the argumentation framework $\mathsf{AF}(D)$ if, and only if (a) $A \in JArgs^D$ and (b) $D \vdash +\partial\phi$;*
- *an argument A and its conclusion ϕ are rejected w.r.t. the argument framework $\mathsf{AF}(D)$ is, and only if (a) $A \in RArgs^D$ and (b) and (b) $D \vdash -\partial\phi$.*

3 Stable Normative Explanations

We define the idea of *normative explanation* for ϕ, which is a normative decision or any piece of normative knowledge that justifies ϕ and that is minimal [11,12,19].

Definition 10 (Normative explanation). *Let $D = (F, R, >)$ be an argumentation theory and $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D. The set* arg $\subseteq \mathscr{A}$ *is a* normative explanation $\mathsf{Expl}(\phi, \mathsf{AF}(D))$ *in $\mathsf{AF}(D)$ for ϕ iff*

- *A* $\in$ arg *is an argument for ϕ and A is justified w.r.t. $\mathsf{AF}(D)$;*
- arg *is a minimal set in $\mathsf{AF}(D)$ such that A is acceptable w.r.t to* arg.

Example 1. Suppose the law forbids engaging in credit activities without a credit license. Such activities are permitted for a person acting on behalf of another person (the principal), when the person is an employee of the principal, and the principal holds a

credit license. Some conditions are specified under which a person can be banned for credit activities. For example, a person is banned if she becomes insolvent.

$$\begin{aligned} R = \{&s_1 : \Rightarrow \neg \mathit{creditActivity},\\ &s_2 : \mathit{creditLicense} \Rightarrow \mathit{creditActivity},\\ &s_3 : \mathit{actsOnBehalfPrincipal}, \mathit{principalCreditLicense} \Rightarrow \mathit{creditActivity},\\ &s_4 : \mathit{banned} \Rightarrow \neg \mathit{creditActivity},\\ &s_5 : \mathit{insolvent} \Rightarrow \mathit{banned}\}\\ >= \{&\langle s_2, > s_1\rangle, \langle s_3 > s_1\rangle, \langle s_4 > s_3\rangle, \langle s_4 > s_2\rangle\}. \end{aligned}$$

Assume an argumentation theory $D = (F, R, >)$ where $F = \{\mathit{insolvent}, \mathit{creditLicense}\}$. Then, $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ is as follows:

$$\begin{aligned} \mathscr{A} = \{&A_1 : \Rightarrow_F \mathit{insolvent}, \quad A_2 : \Rightarrow_F \mathit{creditLicense},\\ &A_3 : A_1 \Rightarrow_{s_5} \mathit{banned}, \quad A_4 : A_3 \Rightarrow_{s_4} \neg \mathit{creditActivity},\\ &A_5 : A_2 \Rightarrow_{s_2} \mathit{creditActivity}\}\\ \gg = \{&\langle A_4, A_5\rangle\rangle\}. \end{aligned}$$

It is easy to see that $\{A_1, A_4\} = \mathsf{Expl}(\neg \mathit{creditActivity}, \mathsf{AF}(D))$.

As discussed in [11,12], an explanation for a given normative conclusion ϕ is stable when adding new elements to that explanation does not affect its power to explain ϕ.

The following definition thus elaborates the ideas of [12] for the argumentation setting of Section 2.2.

Definition 11. *Let R a finite set of rules. We define the set of literals* $\mathrm{Lit}(R)$ *as* $\{\phi, \sim\phi \mid \forall r \in R : \phi \in A(r)$ *or* $\sim\phi \in A(r), R[\phi] \cup R[\sim\phi] = \emptyset\}$.

We write arg_R *to denote the set of all possible arguments that can be built from R and any finite set F of facts such that $F \subseteq \mathrm{Lit}(R)$.*

Definition 12 (Stable Normative Explanation). *Let* $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ *be an argumentation framework determined by the argumentation theory $D = (F, R, >)$. We say that* $\mathsf{arg} = \mathsf{Expl}(\phi, \mathsf{AF}(D))$ *is a* stable normative explanation for ϕ in $\mathsf{AF}(D)$ *iff for all* $\mathsf{AF}(D') = (\mathscr{A}', \gg')$ *where* $D' = (F', R, >)$ *s.t.* $F \subseteq F' \subseteq \mathrm{Lit}(R)$*, we have that* $\mathsf{arg} = \mathsf{Expl}(\phi, \mathsf{AF}(D'))$.

Example 2. Let us consider the argumentation framework $\mathsf{AF}(D)$ in Example 1. Then, $\{A_1, A_4\}$ is stable normative explanation for $\neg \mathit{creditActivity}$ in $\mathsf{AF}(D)$, whereas, e.g., $\{A_2, A_5\}$ is not a stable normative explanation for *creditActivity*.

On the basis of Section 2.2 and Theorem 2 it is easy to verify that the computational results from [11,12] hold also in this case (the proofs are similar and are omitted):

Theorem 3. *Given an argumentation framework* $\mathsf{AF}(D)$ *and a normative explanation, (a) the problem of determining if the explanation is stable is co-NP-complete and (b) the problem of determining if the explanation is not stable is NP-complete.*

4 From Argumentation to Deontic Logic

Let us now show how to move from an argumentation setting to deontic logic.

Intuition 2 *Let $D = (F, R, >)$ be any argumentation theory and $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D. The relation between argumentation and deontic logic is based on the following intution:*

$$D \vdash +\partial\phi \text{ if and only if } \mathscr{M}, w \models \mathbf{OBL}\phi \text{ for some world } w \text{ in some model } \mathscr{M}.$$

The idea behind it—the construction of a canonical model for a defeasible theory—was proposed in [15] for a multi-modal variant of DL where modal operators (including obligations) were explicitly added in the language and proof theory. However, building an argumentation semantics for that formalism is particularly hard, as shown in [16,17].

We avoid those complexities and elaborate on the approach of [15] by constructively stating that defeasible provability of any ϕ corresponds to the obligatoriness of ϕ, and—if **PERM** is the dual of **OBL**—the non-provability of ϕ means that $\sim\phi$ is permitted.

4.1 Deontic Logic and Semantics

Let us define our modal logic language and system.

Definition 13 (Modal language and logic). *Let* Lit *be the set of literals of our language $\mathscr{L}$. The language $\mathscr{L}$*(Lit) *of $\mathbf{E}_{\mathscr{L}}$ is defined as follows:*

$$p ::= l \mid \neg p \mid \mathbf{OBL}\phi \mid \mathbf{PERM}\phi,$$

where l ranges over PROP *and ϕ ranges over* Lit.

The logical system $\mathbf{E}_{\mathscr{L}}$ is based on $\mathscr{L}$(Lit) *and is closed under logical equivalence.*

Proposition 1. *The system $\mathbf{E}_{\mathscr{L}}$ is a fragment of system* **E** *[8].*

Given Proposition 1, we use neighbourhood semantics. However, we have to identify a proper subclass of frames and models.

Let us first recall standard neighbourhood semantics.

Definition 14 (Frames and models). *A* neighbourhood frame *$\mathscr{F}$ is a structure $\langle W, \mathscr{N} \rangle$ where W is a non-empty set of possible worlds and $\mathscr{N}$ is a function $W \mapsto 2^{2^W}$.*

A neighbourhood model *$\mathscr{M}$ is obtained by adding an evaluation function v :* PROP $\mapsto 2^W$ *to a neighbourhood frame.*

Definition 15 (Truth in a model). *Let $\mathscr{M}$ be a model $\langle W, \mathscr{N}, V \rangle$ and $w \in W$. The truth of any formula p in $\mathscr{M}$ is defined inductively as follows:*

1. *standard valuation conditions for the boolean connectives;*
2. *$\mathscr{M}, w \models \mathbf{OBL}\phi$ iff $||\phi|| \in \mathscr{N}(w)$,*
3. *$\mathscr{M}, w \models \mathbf{PERM}\phi$ iff $W - ||\phi|| \notin \mathscr{N}(w)$.*

A formula p is *true at a world* in a model iff $\mathscr{M},w \models p$; *true in a model* $\mathscr{M}$, written $\mathscr{M} \models p$ iff for all worlds $w \in W$, $\mathscr{M},w \models p$; *valid in a frame* $\mathscr{F}$, written $\mathscr{F} \models p$ iff it is true in all models based on that frame; *valid in a class* $\mathscr{C}$ *of frames*, written $\mathscr{C} \models p$, iff it is valid in all frames in the class. Analogously, an inference rule $p_1, \ldots p_n \Rightarrow q$ (where $p_1, \ldots p_n$ are the premises and q the conclusion) is valid in a class $\mathscr{C}$ of frames iff, for any $\mathscr{F} \in \mathscr{C}$, if $\mathscr{F} \models p_1, \ldots, \mathscr{F} \models p_n$ then $\mathscr{F} \models q$.

In order to introduce a semantics for our fragment, the following is needed.

Definition 16. *Let $D = (F, R, >)$ be any argumentation theory, $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D, and* $\mathrm{Lit}(D)$ *be the set of literals occurring in D. The D-*extension $E(D)$ *of a theory D is the smallest set of literals such that, for all* $\phi \in \mathrm{Lit}(D)$*:*

1. $\phi \in E(D)$ *iff* ϕ *is justified w.r.t.* $\mathsf{AF}(D)$*,*
2. $\sim\phi \in E(D)$ *iff* ϕ *is not justified w.r.t.* $\mathsf{AF}(D)$*.*

Definition 17. *Let L be a consistent set of literals. A* defeasible rule theory *is a structure $D = (R, >)$. The D-*extension of *L is the extension of the argumentation theory $(L, R, >)$; we denote it with $E_L(D)$.*

Definition 18 (Dependency graph). *Let D be any argumentation theory and* $\mathrm{Lit}(D)$ *be literals occurring in D. The* dependency graph *of D is the directed graph (V, A) where:*

- $V = \{p \mid p \in \mathrm{PROP}, \{p, \neg p\} \cap \mathrm{Lit}(D) \neq \emptyset\}$*;*
- *A is the set such that $(n, m) \in A$ iff*
  - $n = \phi$ *and* $\exists r \in R[\phi] \cup R[\sim\phi]$*;*
  - $m = \psi$ *and* $\exists r \in R[\psi] \cup R[\sim\psi]$ *such that* $\{n, \sim n\} \cap A(r) \neq \emptyset$*.*

Proposition 2. *Let L be a set of literals, $D = (R, >)$ be a defeasible rule theory such that the transitive closure of $>$ is acyclic and $D' = (L, R, >)$ be the corresponding argumentation theory such that the dependency graph of D' is acyclic. Then, the D-extension of L is consistent iff L is consistent.*

Proof. The result is based on the proofs of Proposition 3.3 of [3], and Theorem 2 of [14] (see [7]). The proof shows that, given the fact that $>$ and the dependency graph of D' are acyclic (which means that we do not have loops in the rules), then, if $D \mid\sim \phi$ and $D \mid\sim \sim\phi$ then $\phi, \sim\phi \in L$, which contradicts the assumption that L is consistent.

The definition of an appropriate structure considers an argumentation theory D: (a) we add as worlds all D-extensions without the empty set, while (b) to construct neighbourhoods for each world, we build an S^r relationship between possible worlds based on information in the rule r for each rule in D and ensure that the rule can actually be applied, and put together all S^r relations.

Definition 19 (Neighbourhood *D*-frame, neighbourhood *D*-model, and truth). *Let $D = (F, R, >)$ be an argumentation theory such that the transitive closure of $>$ is acyclic and the dependency graph of D is acyclic. A* neighbourhood *D-*frame *is a structure $\langle W, \mathscr{N} \rangle$ where*

- $W = \{w \,|\, w \in (2^{E(D)} - \{\emptyset\})\}$;
- $\mathscr{N}$ *is a function with signature* $W \mapsto 2^{2^W}$ *defined as follows:*
  - xS_jy *iff* $\exists r \in R$ *such that* $A(r) \subseteq x$ *and* $C(r) \in y$
  - $\forall s \in R[\sim C(r)]$ *either*
    1. $\exists a \in A(s), a \notin x$; *or*
    2. $\exists t \in R[C(r)]$ *such that* $t > s$, $A(t) \subseteq x$
  - $S_j(w) = \{x \in W : wS_jx\}$
  - $\mathscr{S}_j(w) = \bigcup_{C(r_k)=C(r_j)} S_k(w)$
  - $\mathscr{N}(w) = \{\mathscr{S}_j(w)\}_{r_j \in R}$.

A neighbourhood *D*-model $\mathscr{M}$ *is obtained by adding an evaluation function* $v : \text{PROP} \mapsto 2^W$ *to a neighbourhood D-frame such that, for any* $p \in \text{PROP}$, $v(p) = \{w \,|\, p \in w\}$.

Proposition 3. *Let* $C_{\mathscr{F}}$, $C_{\mathscr{M}}$, $C_{\mathscr{F}_D}$ *and* $C_{\mathscr{M}_D}$ *be, respectively, the classes of neighbourhood frames and models, and the classes of neighbourhood D-frames and D-models. Then,* $C_{\mathscr{F}_D} \subset C_{\mathscr{F}}$ *and* $C_{\mathscr{M}_D} \subset C_{\mathscr{M}}$.

4.2 Completeness

To build canonical structures from an argumentation framework, we use defeasible rule theories by following Intuition 2 and Definitions 16 and 19. The construction considers all possible defeasible rule theories and, for each of them, all possible maximal consistent sets of facts that can be generated. In a nutshell, the procedure runs as follows:

1. **Considering all defeasible rule theories of the language.** Given the language $\mathscr{L}$, the set of all defeasible rule theories is $\mathscr{D}$.
2. **Constructing worlds.** For each defeasible rule theory $D \in \mathscr{D}$, add as worlds all maximal consistent sets of formulae containing all D-extensions of each $L \in 2^{\text{Lit}(D)}$) plus the negation of all literals that do not occur in D.
3. **Constructing neighbourhoods for each world.** Proceed as in Definition 19.

Definition 20 ($\mathscr{L}$-maximality). *A set* w *is* $\mathscr{L}(\text{Lit})$*-maximal iff for any formula* p *of* $\mathscr{L}(\text{Lit})$, *either* $p \in w$, *or* $\neg p \in w$.

Lemma 1 (Lindenbaum's Lemma). *Let D any defeasible rule theory. Any consistent set* $w_{E_L(D)}$ *of formulae in the language* $\mathscr{L}(\text{Lit})$ *consisting of a D-extension of any L can be extended to a consistent* $\mathscr{L}(\text{Lit})$*-maximal set* $w^+_{E_L(D)}$.

Proof. Let $p_1, p_2, \ldots$ be an enumeration of all the possible formulae in $\mathscr{L}(\text{Lit})$.

- $w_0 := w_{E_L(D)}$;
- $w_{n+1} := w_n \cup \{p_n\}$ if its closure under the axioms and rules of a given logic S is consistent, $w_n \cup \{\neg p_n\}$ otherwise;
- $w^+_{E_L(D)} := \bigcup_{n \geq 0} w_n$.

Definition 21 (Canonical neighbourhood *D*-model). *Given the language* $\mathscr{L}$, *let* $\mathscr{D}$ *be the set of all defeasible rule theories that can be obtained from* $\mathscr{L}$. *For all* $D_i = (R_i, >_i) \in \mathscr{D}$ *the* canonical neighbourhood model *is the structure* $\mathscr{M}_{\mathscr{D}} = (W, \mathscr{N}, v)$ *where*

- $W = \bigcup_{\forall D_i \in \mathscr{D}} W_i$ *where* $W_i = \{w_L \,|\, \forall L \in 2^{\mathrm{Lit}(D_i)}, w_L = w^+_{E_L(D)}\}$.
- $\mathscr{N}$ *is a function with signature* $W \mapsto 2^{2^W}$ *defined as follows:*
  - $xS^i_j y$ *where* $\mathbf{OBL}\phi \in x$ *iff* $\exists r \in R_i$ *such that* $C(r) = \phi$, $A(r) \subseteq x$ *and* $C(r) \in y$ *where* $x, y \in W_i$;
  - $\forall s \in R_i[\sim C(r)]$ *either*
    1. $\exists a \in A(s), a \notin x$; *or*
    2. $\exists t \in R_i[C(r)]$ *such that* $t > s$, $A(t) \subseteq x$
  - $S^i_j(w) = \{x \in W_i : wS^i_j x\}$,
  - $\mathscr{S}^i_j(w) = \bigcup_{C(r_k)=C(r_j)} S^i_k(w)$,
  - $\mathscr{N}(w) = \{\mathscr{S}^i_j(w)\}_{r_j \in R_i}$;
- *for each* $\phi \in \mathrm{Lit}$ *and any* $w \in W$, v *is an evaluation function such that* $w \in v(\phi)$ *iff* $\phi \in w$, *and* $w \notin v(\phi)$ *iff* $\sim\phi \in w$.

Lemma 2 (Truth Lemma). *If* $\mathscr{M} = (W, \mathscr{N}, v)$ *is canonical for* S, *where* $\mathsf{S} \supseteq \mathsf{E}_{\mathscr{L}}$, *then for any* $w \in W$ *and for any formula* p, $p \in w$ *iff* $\mathscr{M}, w \models \phi$.

Proof. The proof is by induction on the length of an expression p. We consider only some relevant cases.

Assume p is a literal ϕ. If $\phi \in w$, by the semantic evaluation clause it holds that $\mathscr{M}, w \models \phi$. For the opposite direction, assume that $\mathscr{M}, w \models \phi$, by construction $\phi \in w$.

If, on the other hand, p has the forms $\mathbf{OBL}\phi$ and $\mathbf{PERM}\phi$, and $p \in w$, then, by construction (respectively), $\|\phi\| \in \mathscr{N}(w)$ and $W - \|\phi\| \notin \mathscr{N}(w)$. By definition $\mathscr{M}, w \models \mathbf{OBL}\phi$ and $\mathscr{M}, w \models \mathbf{PERM}\phi$, respectively. Conversely, if $\mathscr{M}, w \models \mathbf{OBL}\phi$ and $\mathscr{M}, w \models \mathbf{PERM}\phi$, then $\|\phi\| \in \mathscr{N}(w)$ and $W - \|\phi\| \notin \mathscr{N}(w)$, and by construction of $\mathscr{N}$, $\mathbf{OBL}\phi \in w$ and $\mathbf{PERM}\phi \in w$.

The canonical model exists, it is not empty, and it is a neighbouhood D-model. Consider any formula $p \notin \mathsf{S}$ such that $\mathsf{S} \supseteq \mathbf{E}_{\mathscr{L}}$; $\{\neg p\}$ is consistent and it can be extended to a maximal set w such that for some canonical model, $w \in W$. By Lemma 2, $w \not\models p$.

Corollary 1 (Completeness of $\mathbf{E}_{\mathscr{L}}$). *The systems* $\mathbf{E}_{\mathscr{L}}$ *is sound and complete with respect to the class of neighbourhood D-frames.*

Corollary 2. *Let* $\mathscr{M}$ *be any neighbourhood D-model. Then*

1. $\mathscr{M} \models \mathbf{OBL}\phi$ *iff there exists an argumentation theory* $D = (F, R, >)$ *such that* ϕ *is justified w.r.t.* $\mathsf{AF}(D)$;
2. $\mathscr{M} \models \mathbf{PERM}\phi$ *iff there exists an argumentation theory* $D = (F, R, >)$ *such that* ϕ *is not justified w.r.t.* $\mathsf{AF}(D)$.

5 Stable Explanations in Neighbourhood Semantics

The definition of normative explanation of Section 3 can be appropriately captured in our deontic logic setting. First of all, we have to formulate the modal version of an argument.

Proposition 4 (Neighbourhood D-model for an argument). *Let $D = (F, R, >)$ be an argumentation theory, $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D, and $\mathscr{M}_D = (W, \mathscr{N}, v)$ be the corresponding neighbourhood D-model. Argument $A \in \mathscr{A}$, where $\mathsf{Conc}(A) = \psi$, is justified w.r.t. $\mathsf{AF}(D)$ iff, if $\mathsf{Sub}(A) = \{A_x \mid \forall x \in \{1, \ldots, s \mid s \geq 1\}, \mathsf{Conc}(A_x) = \phi_{i_x}\}$, then the following condition holds in $\mathscr{M}_D$:*

$$\exists w_1 \ldots \exists w_s \in W \begin{cases} \forall i_2 \in \{1_2, \ldots n_2\}, (\forall i_1 \in \{1_1, \ldots n_1\}, ||\phi_{i_1}|| \in \mathscr{N}(w_1) \Rightarrow \\ \Rightarrow ||\phi_{i_2}|| \in \mathscr{N}(w_1)) \\ \& \\ \forall i_3 \in \{1_3, \ldots n_3\}, (\forall i_2 \in \{1_2, \ldots n_2\}, w_2 \in ||\phi_{i_1}|| \Rightarrow \\ \Rightarrow ||\phi_{i_2}|| \in \mathscr{N}(w_2)) \\ \& \\ \vdots \\ \& \\ \forall i_s \in \{1_s, \ldots n_s\}, (\forall i_{s-1} \in \{1_{s-1}, \ldots n_{s-1}\}, w_{s-1} \in ||\phi_{i_{s-1}}|| \Rightarrow \\ \Rightarrow ||\phi_{i_s}|| \in \mathscr{N}(w_{s-1})) \\ \& \\ \forall i_s \in \{1_s, \ldots n_s\}, (w_s \in ||\phi_{i_s}|| \Rightarrow ||\psi|| \in \mathscr{N}(w_s)) \end{cases}$$

The model $\mathscr{M}_D$ is called a neighbourhood D-model for A.

Proof. ($\Rightarrow$) The argument A has either the form $\Rightarrow_F \psi$ or the form $A_{1_s}, \ldots, A_{n_s} \Rightarrow_r \psi$, where A_{i_s}, $1_s \leq i_s \leq n_s$, is an argument constructed from D, and $r : \mathsf{Conc}(A_{1_s}), \ldots, \mathsf{Conc}(A_{n_s}) \Rightarrow \psi$ is a rule in R.

Case (1). If $A :\Rightarrow_F \psi$, by construction of $\mathscr{M}_D$ (see Definition 19) $s = 1$ and there is a world w_s such that $||\psi|| \in \mathscr{N}(w_s)$.

Case (2). If $A_{1_s}, \ldots, A_{n_s} \Rightarrow_{r_s} \psi$, then, by construction of $\mathscr{M}_D$, there a world w_s such that for each $\mathsf{Conc}(A_{i_s})$, $w_s \in ||\mathsf{Conc}(A_{i_s})||$ and $||\psi|| \in \mathscr{N}(w_s)$. Consider now each A_{i_s}, having the form $A_{1_{s-1}}, \ldots, A_{n_{s-1}} \Rightarrow_{r_{s-1}} \mathsf{Conc}(A_{i_s})$, which in turn can fall within Case (1) or Case (2). Suppose, for example, that each $A_{i_{s-1}}$ falls within Case (2). Then, by construction of $\mathscr{M}_D$, there a world w_{s-1} such that for each $\mathsf{Conc}(A_{i_{s-1}})$, $w_{s-1} \in ||\mathsf{Conc}(A_{i_{s-1}})||$ and $||\mathsf{Conc}(A_{i_s})|| \in \mathscr{N}(w_{s-1})$. Similarly, for the other cases.

Since, A is finite, it means that there are sub-arguments having the form $A_{1_1}, \ldots, A_{n_1} \Rightarrow_{r_1} \phi_{i_2}$ to which we can develop a similar argument.
($\Leftarrow$) The proof for this direction runs similarly as the one for the case ($\Rightarrow$).

The concept of normative explanation directly follows from Proposition 4.

Proposition 5 (Neighbourhood D-model for a normative explanation). *Let $D = (F, R, >)$ be an argumentation theory, $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D, and $\mathscr{M}_D = (W, \mathscr{N}, v)$ be the corresponding neighbourhood D-model.*

If $\mathsf{Expl}(\psi, \mathsf{AF}(D)) = \{A_1, \ldots, A_n\}$ then $\mathscr{M}_D$ is neighbourhood D-model for each argument A_k, $1 \leq k \leq n$.

The model $\mathscr{M}_D$ is called a neighbourhood D-model for $\mathsf{Expl}(\psi, \mathsf{AF}(D))$.

We can semantically isolate the arguments in a normative explanation by using Proposition 4 as well as by resorting to the notion of generated sub-model [5,23].

Definition 22 (Generated submodel [5,23]). *Let $\mathscr{M} = (W, \mathscr{N}, v)$ be any neighbourhood model. A generated submodel $\mathscr{M}_X = (X, \mathscr{N}_X, v_X)$ of $\mathscr{M}$ is neighbourhood model where $X \subseteq W$, $\forall Y \subseteq W, \forall w \in X, Y \in \mathscr{N}(w) \Leftrightarrow Y \cap X \in \mathscr{N}_X(w)$.*

Proposition 6 (Generated D-submodel for a normative explanation). *Let $D = (F, R, >)$ be an argumentation theory, $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D, $\mathscr{X} = \mathsf{Expl}(\psi, \mathsf{AF}(D))$, $\mathscr{M}_D = (W, \mathscr{N}, v)$ be a neighbourhood D-model for $\mathscr{X}$, and $\mathscr{M}_{D_\mathscr{X}} = (W_\mathscr{X}, \mathscr{N}_\mathscr{X}, v_\mathscr{X})$ be a generated submodel of $\mathscr{M}_D$.*

$\mathscr{X} = \{A_1, \ldots, A_n\}$ iff $W_\mathscr{X} = W - X$ where

$$X = \{w \mid w \in W, \forall \phi \in w : \phi \in F \,\&\, A_x \in \mathscr{A}, A_x \notin \mathscr{X} \text{ and } A_x :\Rightarrow_F \phi\}$$

The model $\mathscr{M}_{D_\mathscr{X}}$ is called the generated D-submodel for $\mathscr{X}$.

Proof (Skecth). The model $\mathscr{M}_{D_\mathscr{X}}$ is the generated submodel obtained by isolating in $\mathscr{M}_D$ precisely those worlds, and only those worlds in which the factual literals and factual arguments are those which are needed in $\mathscr{X}$. Hence, (1) if we consider them, all arguments in $\mathscr{X}$ are justified; (2) if we consider only the arguments in $\mathscr{X}$, then, by construction (see Proposition 4), $W_\mathscr{X}$ contains all worlds in W except those in which facts not needed in $\mathscr{X}$ are the case. Notice that, since $\mathscr{M}_{D_\mathscr{X}}$ is a generated submodel, the truth values of modal formulae are preserved.

The semantic reconstruction of stable normative explanation thus trivially follows.

Corollary 3 (Stable normative explanation in neighbourhood D-models). *Let $D = (F, R, >)$ be an argumentation theory and $\mathsf{AF}(D) = (\mathscr{A}, \gg)$ be the argumentation framework determined by D.*

If $\mathscr{X} = \mathsf{Expl}(\psi, \mathsf{AF}(D)) = \{A_1, \ldots, A_n\}$ is a stable normative explanation for ψ in $\mathsf{AF}(D)$ and $D^+ = (F^+, R, >)$ is the argumentation theory where $F^+ = \{\phi \mid \forall r \in R : \phi \in A(r)$ and $R[\phi] \cup R[\sim\phi] = \emptyset\}$, then $\mathsf{Expl}(\psi, \mathsf{AF}(D^+))$, and $\mathscr{M}_{D_\mathscr{X}} = \mathscr{M}_{D^+_\mathscr{X}}$ such that $\mathscr{M}_{D_\mathscr{X}}$ and $\mathscr{M}_{D^+_\mathscr{X}}$ are, respectively, the generated D-submodel and generated D^+-submodel for $\mathscr{X}$.

In other words, a stable explanation considers a neighbourhood model where all possibile facts of a theory D are the case and requires that in such a model the conclusion ψ is still justified.

6 Summary

In this paper we investigated the concept of stable normative explanation in argumentation, which was elsewhere introduced in Defeasible Logic using proof-theoretic methods. Then we have devised in a deontic logic setting a new method to construct appropriate neighborhood models from argumentation frameworks and we have characterised accordingly the notion of stable normative explanation. The problem of determining a stable normative explanation for a certain legal conclusion means to identify a set of facts, obligations, permissions, and other normative inputs able to ensure that such a

conclusion continues to hold when new facts are added to a case. This notion is interesting from a logical point of view—think about the classical idea of inference to the best explanation—and we believe it can also pave the way to develop symbolic models for XAI when applied to the law.

The idea of stability, since it requires to consider adding new inputs, can be reexamined through the revision of the given argumentation theory. Formally, given an initial argumentation theory D_{init}, the revised theory D, and the target conclusion ϕ, we could formally define change operations as follows:

Expansion: from $D_{init} \nvdash \phi$ to $D \vdash \phi$.
Contraction: from $D_{init} \vdash \phi$ to $D \nvdash \phi$.
Revision: from $D_{init} \vdash \phi$ to $D \vdash \sim\phi$.

How such an intuition can be fully exploited in the context of the current research is left to future research.

References

1. Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K.V., Hoos, H.H., Hung, H., Jonker, C.M., Monz, C., Neerincx, M.A., Oliehoek, F.A., Prakken, H., Schlobach, S., van der Gaag, L.C., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., Welling, M.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (2020). https://doi.org/10.1109/MC.2020.2996587, `https://doi.org/10.1109/MC.2020.2996587`
2. Alexy, R.: A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification. Clarendon (1989)
3. Antoniou, G., Billington, D., Governatori, G., Maher, M.: Representation results for defeasible logic. ACM Transactions on Computational Logic **2**(2), 255–287 (2001). https://doi.org/10.1145/371316.371517
4. Atkinson, K., Bench-Capon, T., Bollegala, D.: Explanation in ai and law: Past, present and future. Artificial Intelligence **289**, 103387 (2020). https://doi.org/https://doi.org/10.1016/j.artint.2020.103387, `https://www.sciencedirect.com/science/article/pii/S0004370220301375`
5. van Benthem, J., Pacuit, E.: Dynamic logics of evidence-based beliefs. Studia Logica: An International Journal for Symbolic Logic **99**(1/3), 61–92 (2011), `http://www.jstor.org/stable/41475196`
6. Bex, F., Prakken, H.: On the relevance of algorithmic decision predictors for judicial decision making. In: Maranhão, J., Wyner, A.Z. (eds.) ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021. pp. 175–179. ACM (2021). https://doi.org/10.1145/3462757.3466069, `https://doi.org/10.1145/3462757.3466069`
7. Billington, D.: Defeasible logic is stable. J. Log. Comput. **3**(4), 379–400 (1993). https://doi.org/10.1093/logcom/3.4.379, `https://doi.org/10.1093/logcom/3.4.379`
8. Chellas, B.F.: Modal Logic, An Introduction. Cambridge University Press (1980)
9. Governatori, G., Maher, M.J.: An argumentation-theoretic characterization of defeasible logic. In: Horn, W. (ed.) ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, Germany, August 20-25, 2000. pp. 469–473. IOS Press (2000)

10. Governatori, G., Maher, M.J., Antoniou, G., Billington, D.: Argumentation semantics for defeasible logic. J. Log. Comput. **14**(5), 675–702 (2004)
11. Governatori, G., Olivieri, F., Rotolo, A., Cristani, M.: Inference to the stable explanations. In: LPNMR 2022. pp. 245–258. Springer, Cham (2022)
12. Governatori, G., Olivieri, F., Rotolo, A., Cristani, M.: Stable normative explanations. In: Francesconi, E., Borges, G., Sorge, C. (eds.) Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022. Frontiers in Artificial Intelligence and Applications, vol. 362, pp. 43–52. IOS Press (2022). https://doi.org/10.3233/FAIA220447, https://doi.org/10.3233/FAIA220447
13. Governatori, G., Padmanabhan, V., Rotolo, A., Sattar, A.: A defeasible logic for modelling policy-based intentions and motivational attitudes. Log. J. IGPL **17**(3), 227–265 (2009). https://doi.org/10.1093/jigpal/jzp006, https://doi.org/10.1093/jigpal/jzp006
14. Governatori, G., Rotolo, A.: A computational framework for institutional agency. Artif. Intell. Law **16**(1), 25–52 (2008). https://doi.org/10.1007/s10506-007-9056-y, https://doi.org/10.1007/s10506-007-9056-y
15. Governatori, G., Rotolo, A., Calardo, E.: Possible world semantics for defeasible deontic logic. In: Ågotnes, T., Broersen, J.M., Elgesem, D. (eds.) Deontic Logic in Computer Science - 11th International Conference, DEON 2012, Bergen, Norway, July 16-18, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7393, pp. 46–60. Springer (2012). https://doi.org/10.1007/978-3-642-31570-1_4, https://doi.org/10.1007/978-3-642-31570-1_4
16. Governatori, G., Rotolo, A., Riveret, R.: A deontic argumentation framework based on deontic defeasible logic. In: Miller, T., Oren, N., Sakurai, Y., Noda, I., Savarimuthu, B.T.R., Son, T.C. (eds.) PRIMA 2018: Principles and Practice of Multi-Agent Systems - 21st International Conference, Tokyo, Japan, October 29 - November 2, 2018, Proceedings. Lecture Notes in Computer Science, vol. 11224, pp. 484–492. Springer (2018). https://doi.org/10.1007/978-3-030-03098-8_33, https://doi.org/10.1007/978-3-030-03098-8_33
17. Governatori, G., Rotolo, A., Riveret, R., Villata, S.: Modelling dialogues for optimal legislation. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019. pp. 229–233. ACM (2019). https://doi.org/10.1145/3322640.3326731, https://doi.org/10.1145/3322640.3326731
18. Liao, B., van der Torre, L.: Explanation semantics for abstract argumentation. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020. Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 271–282. IOS Press (2020). https://doi.org/10.3233/FAIA200511, https://doi.org/10.3233/FAIA200511
19. Liu, X., Lorini, E., Rotolo, A., Sartor, G.: Modelling and explaining legal case-based reasoners through classifiers. In: Francesconi, E., Borges, G., Sorge, C. (eds.) Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022. Frontiers in Artificial Intelligence and Applications, vol. 362, pp. 83–92. IOS Press (2022). https://doi.org/10.3233/FAIA220451, https://doi.org/10.3233/FAIA220451
20. Maher, M.J.: Propositional defeasible logic has linear complexity. Theory and Practice of Logic Programming **1**(6), 691–711 (2001)
21. Medvedeva, M., Vols, M., Wieling, M.: Using machine learning to predict decisions of the european court of human rights. Artificial Intelligence and Law **28**(2), 237–266 (2020). https://doi.org/10.1007/s10506-019-09255-y

22. Odekerken, D., Bex, F., Borg, A., Testerink, B.: Approximating stability for applied argument-based inquiry. Intell. Syst. Appl. **16**, 200110 (2022). https://doi.org/10.1016/j.iswa.2022.200110, https://doi.org/10.1016/j.iswa.2022.200110
23. Pacuit, E.: Neighborhood Semantics for Modal Logic. Cham, Switzerland: Springer (2017)
24. Peczenik, A.: On Law and Reason. Kluwer, Dordrecht (1989)
25. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument and Computation **1**(2), 93–124 (2010)
26. Prakken, H., Ratsma, R.: A top-level model of case-based argumentation for explanation: Formalisation and experiments. Argument Comput. **13**(2), 159–194 (2022). https://doi.org/10.3233/AAC-210009, https://doi.org/10.3233/AAC-210009
27. Prakken, H., Sartor, G.: Law and logic: A review from an argumentation perspective. Artif. Intell. **227**, 214–245 (2015). https://doi.org/10.1016/j.artint.2015.06.005, https://doi.org/10.1016/j.artint.2015.06.005
28. Riveret, R., Rotolo, A., Sartor, G.: A deontic argumentation framework towards doctrine reification. FLAP **6**(5), 903–940 (2019), https://collegepublications.co.uk/ifcolog/?00034