




Stable Normative Explanations: From Argumentation to Deontic Logic

Cecilia Di Florio¹ , Antonino Rotolo¹  , Guido Governatori² ,
and Giovanni Sartor^{1,3} 

¹ ALMA AI and Department of Legal Studies, University of Bologna, Bologna, Italy
{cecilia.diflorio2,antonino.rotolo,giovanni.sartor}@unibo.it

² Coorobah QLD 4565, Australia
guido@governatori.net

³ EUI, Fiesole, Italy

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

1 Introduction

Developing explainable AI systems is important in the law since ‘*transparency*’ and ‘*justification*’ of legal decision-making require formalising normative explanations [1, 4, 6, 15]. A normative explanation is an explanation where norms are crucial: in the context of legal decision-making, this means to explain why a legal conclusion ought to be the case on the basis of certain norms and facts [2, 10, 13, 14, 18, 19].

Legal proceedings are adversarial in nature: 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* [9, 10, 16]. 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? How to ensure that the facts presented by a party are ‘resilient’ to the attacks from the opponent? In this paper we adopt [9, 10]’s definition of stability and elaborate it in the argumentation setting of Defeasible Logic [3].

What is the deontic meaning of stable normative explanation as developed in an argumentation setting? In legal argumentation, a typical outcome of judicial decisions are obligations and permissions. In moving to the deontic domain, we notice that deontic argumentation can be developed in various ways [12, 21]. We assume that legal norms are rules having the form $\phi_1, \dots, \phi_n \Rightarrow \psi$ and we follow the intuition that, if

Antonino Rotolo and Giovanni Sartor were partially supported by the Project PE01 “Future AI Research” (FAIR, PNRR, CUP: J33C22002830006).

AF is an argumentation framework where arguments are built using rules, then **OBL** ψ holds in AF iff ψ is justified w.r.t. AF [20]. Once this is done and we have defined the notion of normative explanation, we adapt [11]’s method and show how *this machinery can be reconstructed in neighborhood semantics for classical deontic logics [7] and how the notion of explanation can be semantically characterised.*

The layout of article is as follows¹. Section 2 offers a variant of the idea of argumentation framework based on Defeasible Logic. Section 3 presents the definition of stable normative explanation. Section 4 illustrates how to move from argumentation structures to neighbourhood semantics for deontic logic. Section 5 applies the ideas of Sects. 3 and 4 to semantically reconstruct the concept of stable normative explanation.

2 Background: Logic and Argumentation

Our framework is Defeasible Logic (DL) [3]. The basic language consists of a set Lit of literals. The *complementary* of a literal ϕ is denoted by $\sim\phi$: if ϕ is positive then $\sim\phi$ is $\neg\phi$, if ϕ is negative then $\sim\phi$ is ϕ . Let Lab be a set of labels to represent names of rules. A rule r has the form $r: A(r) \Rightarrow C(r)$, where: (i) $r \in \text{Lab}$ is the unique name of the rule, (ii) $A(r) \subseteq \text{Lit}$ is r ’s (set of) antecedents, (iii) $C(r) = \phi \in \text{Lit}$ is its conclusion. Unlike standard DL, *we only use defeasible rules*, in which, if the premises are the case, then typically the conclusion holds unless we have contrary evidence.

We also use a special type of logical theory in DL:

Definition 1 (Argumentation theory). *An argumentation theory D is a tuple $(F, R, >)$ where (a) $F \subseteq \text{Lit}$ is a finite and consistent set of facts (indisputable statements), (b) R is a finite rule set, and (c) a binary superiority relation over R (which is used to solve rule conflicts). We state that $\forall\phi \in F, R[\phi] \cup R[\sim\phi] = \emptyset$.*

As a convention, $R[\phi]$ denotes the set of all rules in R whose conclusion is ϕ .

A *conclusion* of D is a tagged literal with the following form: $+\partial\phi$ (resp. $-\partial\phi$) means that ϕ is *defeasibly proved* (resp. *defeasibly refuted*) in D , i.e., there is a defeasible proof for ϕ in D (resp. a proof does not exist). A proof P of length n in D is a finite sequence $P(1), P(2), \dots, P(n)$ of tagged literals for which specific proof conditions are defined [3]. $P(1..n)$ denotes the first n steps of P . We present only the positive one for defeasible conclusions.

$+\partial\phi$: If $P(n+1) = +\partial\phi$ then either

- (1) $\phi \in F$, or
- (2.1) $\exists r \in R[\phi]$ s.t. $\forall\psi \in A(r). +\partial\psi \in P(1..n)$ and
- (2.2) $\forall s \in R[\sim\phi]$ either
 - (2.2.1) $\exists\psi \in A(s). -\partial\psi$, or
 - (2.2.2) $\exists t \in R[\phi]$ s.t. $\forall\psi \in A(t). +\partial\psi \in P(1..n)$ and $t > s$.

Argumentation frameworks for DL have been studied in [8]. Here, we present a variant of it, which is based on the above fragment of DL [9, 10].

¹ A full version of this paper with some proofs is here: <http://arxiv.org/abs/2307.05156>.

Definition 2 (Argument). Let $D = (F, R, >)$ be an argumentation theory. An argument A constructed from D has either the form $\Rightarrow_F \phi$ (factual argument), where $\phi \in F$, or the form $A_1, \dots, A_n \Rightarrow_r \phi$ (plain argument), where $1 \leq k \leq n$, and

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

For a given argument A , Conc returns its conclusion, Sub returns all its sub-arguments, and TopRule returns the last rule in the argument.

Any argument A is a tree whose root is labelled by $\text{Conc}(A)$, and for every node x labelled by any ϕ , its children x_1, \dots, x_n are labelled by ϕ_1, \dots, ϕ_n (except its leaves, which can be also labelled by \emptyset) and the arcs are labeled by a rule $r : \phi_1, \dots, \phi_n \Rightarrow \phi$. Arguments of height 1 are called *atomic arguments*; for any argument A , the set of its atomic sub-arguments is denoted by $\text{ASub}(A)$.

The notions of *attack*, *support*, and *undercut* are the standard ones for DL (see [8]). We can now define the argumentation framework.

Definition 3 (Argumentation Framework). Let D be an argumentation theory. The argumentation framework $\text{AF}(D)$ determined by D is (\mathcal{A}, \gg) where \mathcal{A} is the set of all arguments constructed from D , and \gg is the attack relation.

The core of argumentation semantics are the notions of *acceptable* and *rejected argument*. An argument is acceptable with respect to a set of arguments that undercut any attacks. Then, we can define recursively the *extension of an argumentation theory D* and of the corresponding framework $\text{AF}(D)$, which is the set of *justified arguments* w.r.t. $\text{AF}(D)$. The definitions of the set JArgs^D of *justified arguments* and of the set RArgs^D of *rejected arguments* are a fix-point construction. For the details see [8].

Theorem 1. Let D be an argumentation theory and A be an argument in $\text{AF}(D)$ such that $\text{Conc}(A) = \phi$. Then, (a) $A \in \text{JArgs}^D$ iff $D \vdash +\partial\phi$; (b) $A \in \text{RArgs}^D$ iff $D \vdash -\partial\phi$.

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 [9, 10, 14].

Definition 4 (Normative explanation). Let $D = (F, R, >)$ be an argumentation theory and $\text{AF}(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D . The set $\text{arg} \subseteq \mathcal{A}$ is a normative explanation $\text{Expl}(\phi, \text{AF}(D))$ in $\text{AF}(D)$ for ϕ iff

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

Example 1. Consider the following fragment of an argumentation theory:

$$R = \{s_1: \Rightarrow \neg\alpha, s_2: \lambda \Rightarrow \alpha, s_3: \beta, \pi \Rightarrow \alpha, s_4: \delta \Rightarrow \neg\alpha, s_5: \iota \Rightarrow \delta\}$$

$$\Rightarrow = \{\langle s_2, > s_1 \rangle, \langle s_3 > s_1 \rangle, \langle s_4 > s_3 \rangle, \langle s_4 > s_2 \rangle\}.$$

Assume an argumentation theory $D = (F, R, >)$ where $F = \{\iota, \lambda\}$. Then, $\text{AF}(D) = (\mathcal{A}, \gg)$ is as follows:

$$\begin{aligned} \mathcal{A} &= \{A_1: \Rightarrow_F \iota, A_2: \Rightarrow_F \lambda, A_3: A_1 \Rightarrow_{s_5} \delta, A_4: A_3 \Rightarrow_{s_4} \neg\alpha, A_5: A_2 \Rightarrow_{s_2} \alpha\} \\ \gg &= \{\langle A_4, A_5 \rangle\}. \end{aligned}$$

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

An explanation for a normative conclusion ϕ is stable when adding new elements to that explanation does not affect its power to explain ϕ [9, 10].

Definition 5. Let R a finite set of rules. We define the set of literals $\text{Lit}(R)$ as $\{\phi, \sim\phi \mid \forall r \in R: \phi \in A(r) \text{ or } \sim\phi \in A(r), R[\phi] \cup R[\sim\phi] = \emptyset\}$.

Definition 6 (Stable Normative Explanation). Let $\text{AF}(D) = (\mathcal{A}, \gg)$ be an argumentation framework determined by the argumentation theory $D = (F, R, >)$. We say that $\text{arg} = \text{Expl}(\phi, \text{AF}(D))$ is a stable normative explanation for ϕ in $\text{AF}(D)$ iff for all $\text{AF}(D') = (\mathcal{A}', \gg')$ where $D' = (F', R, >)$ s.t. $F \subseteq F' \subseteq \text{Lit}(R)$, we have that $\text{arg} = \text{Expl}(\phi, \text{AF}(D'))$.

Example 2. Let us consider Example 1. Then, $\{A_1, A_4\}$ is stable normative explanation for $\neg\alpha$ in $\text{AF}(D)$, whereas, e.g., $\{A_2, A_5\}$ is not a stable normative explanation for α .

4 From Argumentation to Deontic Logic

To move to deontic logic we follow [11] by stating that defeasible provability (and justification) of any ϕ corresponds to the obligatoriness of ϕ , and—if **PERM** is the dual of **OBL**—the non-provability (and non-justification) of ϕ means that $\sim\phi$ is permitted.

Definition 7 (Modal language and logic). Let Lit be the set of literals of our language \mathcal{L} . The language $\mathcal{L}(\text{Lit})$ of $\mathbf{E}_{\mathcal{L}}$ 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 $\mathbf{E}_{\mathcal{L}}$ is based on $\mathcal{L}(\text{Lit})$ and is closed under logical equivalence.

Proposition 1. The system $\mathbf{E}_{\mathcal{L}}$ is a fragment of system **E** [7].

To introduce an appropriate semantics for our fragment, the following is needed.

Definition 8. Let $D = (F, R, >)$ be any argumentation theory, $\text{AF}(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D , and $\text{Lit}(D)$ 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 $\phi \in \text{Lit}(D)$: (a) $\phi \in E(D)$ iff ϕ is justified w.r.t. $\text{AF}(D)$, (b) $\sim\phi \in E(D)$ iff ϕ is not justified w.r.t. $\text{AF}(D)$.

Definition 9. 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 10 (Dependency graph). Let D be any argumentation theory and $\text{Lit}(D)$ be literals occurring in D . The dependency graph of D is the directed graph (V, A) where:

- $V = \{p \mid p \in \text{PROP}, \{p, \neg p\} \cap \text{Lit}(D) \neq \emptyset\}$;
- 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.

Definition 11. (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. Let $R' = R \cup \{r := \phi \mid \phi \in F\}$. A neighbourhood D -frame is a structure $\langle W, \mathcal{N} \rangle$ where

- $W = \{w \mid w \in (2^{E(D)} - \{\emptyset\})\}$;
- \mathcal{N} is a function with signature $W \mapsto 2^{2^W}$ defined as follows:
 - $xS_j y$ iff $\exists r_j \in R'$ such that $A(r_j) \subseteq x$ and $C(r_j) \in y$
 - $\forall s \in R'[\sim C(r_j)]$ either
 1. $\exists a \in A(s), a \notin x$; or
 2. $\exists t \in R'[C(r_j)]$ such that $t > s, A(t) \subseteq x$
 - $S_j(w) = \{x \in W : wS_j x\}$
 - $\mathcal{S}_j(w) = \bigcup_{C(r_k)=C(r_j)} S_k(w)$
 - $\mathcal{N}(w) = \{\mathcal{S}_j(w)\}_{r_j \in R'}$.

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

To build canonical structures, we consider all possible defeasible rule theories and, for each of them, all possible maximal consistent sets of facts that can be generated.

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

Definition 12. (Canonical neighbourhood D -model). Given the language \mathcal{L} , let \mathcal{D} be the set of all defeasible rule theories that can be obtained from \mathcal{L} . For all $D_i = (R_i, >_i) \in \mathcal{D}$, define $R'_i = R_i \cup \{r := \phi \mid \phi \in L\}$ for each $(L, R_i, >_i)$, $L \in 2^{\text{Lit}(D_i)}$. The canonical neighbourhood model is the structure $\mathcal{M}_{\mathcal{D}} = (W, \mathcal{N}, v)$ where

- $W = \bigcup_{\forall D_i \in \mathcal{D}} W_i$ where $W_i = \{w_L \mid \forall L \in 2^{\text{Lit}(D_i)}, w_L = w_{E_L(D_i)}^+\}$.
- \mathcal{N} is a function with signature $W \mapsto 2^{2^W}$ defined as follows:
 - $xS'_j y$ where **OBL** $\phi \in x$ iff $\exists r_j \in R'_i$ such that $C(r_j) = \phi$, $A(r_j) \subseteq x$ and $C(r_j) \in y$ where $x, y \in W_i$;
 - $\forall s \in R'_i[\sim C(r_j)]$ either

1. $\exists a \in A(s), a \notin x$; or
 2. $\exists t \in R'_i[C(r_j)]$ such that $t > s, A(t) \subseteq x$
- $S_j^i(w) = \{x \in W_i : wS_j^i x\}$,
 - $\mathcal{S}_j^i(w) = \bigcup_{C(r_k)=C(r_j)} S_k^i(w)$,
 - $\mathcal{N}(w) = \{\mathcal{S}_j^i(w)\}_{r_j \in R'_i}$;
- for each $\phi \in \text{Lit}$ 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 $\mathcal{M} = (W, \mathcal{N}, v)$ is canonical for S , where $S \supseteq E_{\mathcal{L}}$, then for any $w \in W$ and for any formula ϕ , $\phi \in w$ iff $\mathcal{M}, w \models \phi$.*

Corollary 1. (Completeness of $E_{\mathcal{L}}$). *The system $E_{\mathcal{L}}$ is sound and complete with respect to the class of neighbourhood D -frames.*

Corollary 2. *Let \mathcal{M} be any neighbourhood D -model. Then (a) $\mathcal{M} \models \text{OBL}\phi$ iff there exists an argumentation theory $D = (F, R, >)$ such that ϕ is justified w.r.t. $\text{AF}(D)$; (b) $\mathcal{M} \models \text{PERM}\phi$ iff there exists an argumentation theory $D = (F, R, >)$ such that $\neg\phi$ is not justified w.r.t. $\text{AF}(D)$.*

5 Stable Explanations in Neighbourhood Semantics

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

Proposition 3 (Neighbourhood D -model for an argument). *Let $D = (F, R, >)$ be an argumentation theory, $\text{AF}(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D , and $\mathcal{M}_D = (W, \mathcal{N}, v)$ be the corresponding neighbourhood D -model. An argument $A \in \mathcal{A}$, where $\text{Conc}(A) = \phi_0$, is justified w.r.t. $\text{AF}(D)$ iff, if h is the height of A and $\mathcal{A} = \{A_x \mid A_x \in \text{ASub}(A), \forall x \in \{(h-1)_1, \dots, (h-1)_m, \dots, 1_1, \dots, 1_p, 0\}, \text{Conc}(A_x) = \phi_x\}$, then the following condition holds in \mathcal{M}_D : if $y \in \{h_1, \dots, h_m, (h-1)_1, \dots, (h-1)_m, \dots, 1_1, \dots, 1_p, 0\}$*

$$\exists w_y \in W \left\{ \begin{array}{l} \forall (h-1)_z \in \{(h-1)_1, \dots, (h-1)_m\}, (\dots (\|\phi_{(h-1)_z}\| \in \mathcal{N}(w_{h_z})) \\ \& \\ \forall (h-2)_z \in \{(h-2)_1, \dots, (h-2)_j\}, (w_{(h-1)_z} \in \|\phi_{(h-1)_z}\| \Rightarrow \\ \Rightarrow \|\phi_{(h-2)_z}\| \in \mathcal{N}(w_{(h-1)_z})) \\ \& \\ \vdots \\ \& \\ \forall 2_z \in \{2_1, \dots, 2_k\}, (w_{2_z} \in \|\phi_{2_z}\| \Rightarrow \|\phi_{1_z}\| \in \mathcal{N}(w_{2_z})) \\ \& \\ \forall 1_z \in \{1_1, \dots, 1_j\}, (w_{1_z} \in \|\phi_{1_z}\| \Rightarrow \|\phi_0\| \in \mathcal{N}(w_{1_z})) \dots \end{array} \right.$$

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

The concept of normative explanation directly follows from Proposition 3.

Proposition 4 (Neighbourhood D -model for a normative explanation). *Let $D = (F, R, >)$ be an argumentation theory, $\text{AF}(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D , and $\mathcal{M}_D = (W, \mathcal{N}, v)$ be the corresponding neighbourhood D -model.*

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

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

We can semantically isolate the arguments in a normative explanation by using Proposition 3 as well as by resorting to the notion of generated sub-model [5, 17].

Definition 13 (Generated submodel [5, 17]). *Let $\mathcal{M} = (W, \mathcal{N}, v)$ be any neighbourhood model. A generated submodel $\mathcal{M}_X = (X, \mathcal{N}_X, v_X)$ of \mathcal{M} is neighbourhood model where $X \subseteq W$, $\forall Y \subseteq W, \forall w \in X, Y \in \mathcal{N}(w) \Leftrightarrow Y \cap X \in \mathcal{N}_X(w)$.*

Proposition 5 (Generated D -submodel for a normative explanation). *Let $D = (F, R, >)$ be an argumentation theory, $\text{AF}(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D , $\mathcal{X} = \text{Expl}(\psi, \text{AF}(D))$, $\mathcal{M}_D = (W, \mathcal{N}, v)$ be a neighbourhood D -model for \mathcal{X} , and $\mathcal{M}_{D_{\mathcal{X}}} = (W_{\mathcal{X}}, \mathcal{N}_{\mathcal{X}}, v_{\mathcal{X}})$ be a generated submodel of \mathcal{M}_D .*

$\mathcal{X} = \{A_1, \dots, A_n\}$ iff $W_{\mathcal{X}} = W - X$ where

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

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

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

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

A stable explanation considers a neighbourhood model where all possible 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. 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.

References

1. Akata, Z., et al.: 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>
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 Trans. Comput. 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. *Artif. Intell.* **289**, 103387 (2020) <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: Int. J. Symbol. Logic* **99**(1/3), 61–92 (2011). <https://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.) *Eighteenth International Conference for Artificial Intelligence and Law, ICAIL 2021, São Paulo Brazil, 21–25 June 2021*, pp. 175–179. ACM (2021). <https://doi.org/10.1145/3462757.3466069>
7. Chellas, B.F.: *Modal Logic: An Introduction*. Cambridge University Press, Cambridge (1980)
8. Governatori, G., Maher, M.J., Antoniou, G., Billington, D.: Argumentation semantics for defeasible logic. *J. Log. Comput.* **14**(5), 675–702 (2004)
9. Governatori, G., Olivieri, F., Rotolo, A., Cristani, M.: Inference to the stable explanations. In: Gottlob, G., Incezan, D., Maratea, M. (eds.) *LPNMR 2022*. LNCS, pp. 245–258. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-15707-3_19
10. 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>
11. Governatori, G., Rotolo, A., Calardo, E.: Possible world semantics for defeasible deontic logic. In: Ágotnes, T., Broersen, J., Elgesem, D. (eds.) *DEON 2012*. LNCS (LNAI), vol. 7393, pp. 46–60. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31570-1_4
12. 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., Cao Son, T. (eds.) *PRIMA 2018*. LNCS (LNAI), vol. 11224, pp. 484–492. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03098-8_33
13. 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, 4–11 September 2020*. *Frontiers in Artificial Intelligence and Applications*, vol. 326, pp. 271–282. IOS Press (2020). <https://doi.org/10.3233/FAIA200511>
14. 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>
15. Medvedeva, M., Vols, M., Wieling, M.: Using machine learning to predict decisions of the European court of human rights. *Artif. Intell. Law* **28**(2), 237–266 (2020). <https://doi.org/10.1007/s10506-019-09255-y>

16. 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>
17. Pacuit, E.: *Neighborhood Semantics for Modal Logic*. Springer, Cham, Switzerland (2017)
18. Peczenik, A.: *On Law and Reason*. Kluwer, Dordrecht (1989)
19. 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>
20. 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>
21. 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>