Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220447

Stable Normative Explanations

Guido GOVERNATORI [a] Francesco OLIVIERI [b],
Antonino ROTOLO [c], Matteo CRISTANI [d]
[a] *Centre for Computational Law, Singapore Management University, Singapore*
[b] *Institute for Integrated and Intelligent Systems, Griffith University, Australia*
[c] *Alma AI, University of Bologna, Italy*
[d] *Department of Computer Science, University of Verona, Italy*

Abstract. Modelling the concept of explanation is a central matter in AI systems, as it provides methods for developing eXplainable AI (XAI). When explanation applies to normative reasoning, XAI aims at promoting normative trust in the decisions of AI systems: in fact, such a trust depends on understanding whether systems predictions correspond to legally compliant scenarios. This paper extends to normative reasoning a work by Governatori *et al.* (2022) on the notion of stable explanations in a non-monotonic setting: when an explanation is stable, it can be used to infer the same normative conclusion independently of other facts that are found afterwards.

Keywords. Defeasible Deontic Logic, Stable Explanation, Symbolic XAI

1. Introduction

The literature on the concept of *explanation* is vast (especially in philosophy; see, among many others, [1,14]), and the AI community is recently paying more and more attention to it due to the development of eXplainable AI (XAI) [12]. The AI&Law community has, in turn, a long tradition in this direction [3], since '*transparency*' and '*justification*' of legal decision-making both require formalising normative explanations.

We propose the novel idea of *stable normative explanation* extending the notion of *stable inference* of [4]. Roughly speaking, 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), but it can also pave the way to develop symbolic models for XAI when applied to the Law (consider, for instance, systems of predictive justice [11]).

Reasoning with legal norms exhibits features that distinguish it from other types of reasoning. For instance, while examining a case, we are limited to (i) the facts presented (and the admissible ones) for that case, and (ii) the norms in force for the time relevant to the case itself, norms that, sometimes, stem from different sources. Given the facts of the case, the proceeding aims at determining what *legal requirements* (expressed as obligations, prohibitions and permissions) hold, and whether such legal requirements have been fulfilled. If more/new facts were presented, the outcome of a case might be quite different or can even be modified; moreover, such additional facts may be themselves the outcome

of some norms establishing other legal requirements (i.e., obligations, prohibitions and permissions), possibly not in the sphere of control of the current proceeding.

Accordingly, one of the major issues is how to ensure a specific outcome for a case, which, in an adversarial setting, can be understood as how to ensure that the facts presented by a party are 'resilient' to the attacks from the opponent. Furthermore, such an issue is not restricted to adversarial situations only. Still, it is relevant even in cases where one party does their due diligence to determine if they comply with a particular piece of legislation (e.g., to identify all requirements that a business must satisfy to legally serve alcohol in an entertainment environment, according to the Queensland liquor licensing and gaming regulations).

Let us ground the above discussion with a concrete scenario. The Australian Spent Conviction discipline governs (1) the conditions under which a conviction is spent, and (2) when it is permitted to withhold information about it (or when it is mandatory to disclose that a conviction has occurred). At the federal level, Spent Conviction is regulated by Part VII C of the Crimes Act 1914; in addition, States and Territories enacted their own legislation and schemas supplementing and complementing the federal one.

Part VII C of the Crimes Act 1914 consists of six divisions. Division 1 gives the terms and definitions for the topics. Divisions 2, 3, and 4 establish the baseline conditions for (i) when a conviction is spent, (ii) when a person is permitted to withhold information about a (spent) conviction, (iii) when a person is required to provide such information, and (iv) when third parties are either permitted or forbidden to disclose information they might have about a (spent) conviction. Division 5 deals with "administrative" aspects (complaints) of improper releases of spent conviction information. Finally, Division 6 specifies the exclusions (exceptions) to Divisions 2 and 3. Among the various provisions, Section 85ZZGB (Exclusion: disclosing information to a person or body) recites:

> Divisions 2 and 3 do not apply in relation to the disclosure of information to a prescribed person or body if:
>
> (a) The person or body is required or permitted by or under a prescribed Commonwealth law, a prescribed State law or a prescribed Territory law, to obtain and deal with information about persons who work, or seek to work, with children; and
>
> (b) The disclosure is for the purpose of the person or body obtaining and dealing with such information in accordance with the prescribed law.

Part VII C of the Crimes Act 1914 clearly demonstrates the abovementioned issues. First of all, examining the baseline conditions given in Divisions 2 and 3 is not sufficient to determine if the information about a spent conviction can be withheld: the exclusions specified in Division 4 must be considered. Moreover, Section 85ZZGB specifies that the conditions set in Division 2 or 3 depend upon deontic aspects ("required" or "permitted") determined by regulatory instruments outside what is specified by Part VII C, which can be assumed as (external deontic) facts of the case.

In this paper, we work with a non-monotonic formalism (Defeasible Deontic Logic) apt to model norms and able to deal with exceptions and deontic concepts. The logic is needed to provide a precise and formal grounding of the problem of *stability* and *stable normative explanation*. We also examine the computational complexity of ensuring stability.

2. Stable Explanations

Finding a normative explanation for a certain normative conclusion l (such as an obligation) means determining as input a piece F of normative information that supports the derivation of l through norms and other rules of the normative system. If F is a stable explanation, then adding new facts to that explanation does not affect its power to explain the normative conclusion. Stable explanations are naturally considered because they are insensitive to input knowledge changes. In other terms, stable explanations are, to some extent, *monotonic*, even when the considered logic is not.

In this investigation, we work on a deontic extension of Defeasible Logic (DL), called Defeasible Deontic Logic (DDL). In DDL, we have three types of elements: (1) facts, (2) rules, and (3) superiority relation $>$. Facts are the input knowledge describing those indisputable things that are true beyond any doubt. Rules are ways to obtain (normative) conclusions that are considered plausible (or typical), whereas the superiority relation is thought of as a means to establish whether one rule for a conclusion might prevail against another rule for the opposite conclusion.

DDL is DL plus the deontic operators O and P, respectively, for obligations and permissions, and the operator $\otimes$ according to which an expression $a \otimes b$ means that a is obligatory, but if such an obligation is violated, then b is obligatory and compensates this violation [8]. In addition to standard rules (which are hereafter referred to as *constitutive rules*, with arrow $\Rightarrow_{\mathsf{C}}$), we have deontic rules, such as

$$\alpha\colon a_1, \ldots, a_n \Rightarrow_{\mathsf{O}} b \otimes c \qquad\qquad \beta\colon \mathsf{O}c, d_1, \ldots, d_m \Rightarrow_{\mathsf{P}} e.$$

If α is applicable (namely, $a_1, \ldots, a_n$ are the case), then we derive $\mathsf{O}b$. Suppose that we know $\neg b$, meaning that $\mathsf{O}b$ is violated. In this case, we derive $\mathsf{O}c$. Accordingly, if we also know that $d_1, \ldots, d_m$ are the case, then we conclude that e is permitted, i.e., that $\mathsf{P}e$.

Two peculiar features of DDL make the idea of normative explanation not obvious:

- The set F of facts may include deontic expressions such as $\mathsf{O}p$, and $\neg\mathsf{P}q$. Such deontic facts encode a *normatively* indisputable input. For example, suppose the set of rules represents norms of the Italian legal system. In that case, we can take $\mathsf{O}b \in F$ as indisputable as grounded on the Italian constitution or because it is imported from European law.
- DDL adopts the concept of rule conversion [9], which amounts here to use non-deontic rules to derive obligations and permissions. Consider the rule $\alpha\colon a \Rightarrow_{\mathsf{C}} b$, and assume we prove $\mathsf{P}a$. We can use α to determine that b is permitted. For example, in football, if the ball passing completely over the goal line between the goal posts and under the crossbar 'counts as' scoring a goal and it is permitted for the ball to pass such a goal line, then we can indeed derive that scoring a goal is permitted.

To illustrate the idea of stable normative explanation, consider the following example.

Example 1. *Suppose the Law forbids engaging in credit activities without a credit license. If you violate this prohibition, the civil penalty is 2,000 penalty units. Furthermore, such activities are permitted for a person acting on behalf of another person (the principal) when the person is an employee or the director of the principal and the principal holds a credit license. Moreover, some conditions are specified under which a person could be banned from credit activities. For example, a person is banned if they become insolvent. Finally, using the equity mobilised by the credit institutions counts as a credit activity.*

γ: $\Rightarrow_{\mathsf{O}} \neg creditActivity \otimes civilPenalty$
δ: $creditLicense \Rightarrow_{\mathsf{P}} creditActivity$
ϵ: $actsOnBehalfPrincipal, principalCreditLicense \Rightarrow_{\mathsf{P}} creditActivity$
ζ: $\mathsf{O}banned \Rightarrow_{\mathsf{O}} \neg creditActivity$
η: $insolvent \Rightarrow_{\mathsf{O}} banned$
θ: $\mathsf{O}pay, \neg pay \Rightarrow_{\mathsf{C}} insolvent$
ι: $equity \Rightarrow_{\mathsf{C}} creditActivity$

where $\delta > \gamma$, $\epsilon > \gamma$, $\zeta > \epsilon$, $\iota > \gamma$, *and* $\iota > \zeta$.

- *The set* {*creditLicence*} *is stable for* P*creditActivity;*
- *The set* {*actsOnBehalfPrincipal*, *principalCreditLicense*} *is not stable for*

- *The set* {P*equity*} *is stable for* P*creditActivity;*
- *The set* {*creditAcvitiy*} *is not stable for the conclusion* O*civilPenalty (if we add, e.g., creditLicence).*

3. Defeasible Deontic Logic

Defeasible Logic [13,2] is a simple, flexible, and efficient rule-based non-monotonic formalism, whose strength lies in its constructive proof theory that allows drawing meaningful conclusions from a (potentially) conflicting and incomplete knowledge base. In non-monotonic systems, more accurate conclusions can be obtained when more pieces of information become available. Many variants of DL have been proposed for the logical modelling of different application areas, especially for legal reasoning (for an overview of the literature, see [10]).

In this research, we focus on the Defeasible Deontic Logic's framework advanced in [5], that allows us to model a large variety of normative concepts, as well as to determine what prescriptive behaviours are in force in a given situation.

We shall now briefly recall the main elements of the logic, and start by defining the language of a defeasible deontic theory. Let PROP be a set of propositional atoms, and Lab be a set of arbitrary labels (the names of the rules). Lower-case Roman letters denote literals, whereas lower-case Greek letters denote rules. Accordingly, PLit = PROP $\cup$ $\{\neg l \mid l \in \text{PROP}\}$ is the set of *plain literals*, the set of *deontic literals* is ModLit = $\{\Box l, \neg\Box l \mid l \in \text{PLit} \wedge \Box \in \{\mathsf{O}, \mathsf{P}\}\}$, and finally, the set of *literals* is Lit = PLit $\cup$ ModLit. The *complement* of a literal l is denoted by $\sim l$: if l is a positive literal p then $\sim l$ is $\neg p$, and if l is a negative literal $\neg p$ then $\sim l$ is p.

Definition 1 (Defeasible Deontic Theory)**.** *A* defeasible deontic theory D *is a tuple* $(F, R, >)$*, where* F *is the set of facts,* R *is the set of rules, and* $>$ *is a binary relation over* R *(called* superiority relation*).*

The set of facts $F \subseteq$ Lit denotes simple pieces of information that are considered to be always true. A theory is meant to represent a normative system, where the rules encode the norms of such a system, and the set of facts corresponds to "*factual information*", or "*given deontic positions*" (as, for instance, indisputable obligations or deontic positions imported from other higher-ranked normative systems). The rules are used to conclude the institutional facts, obligations and permissions that hold given the set of facts.

The set of rules R is finite and contains three *types* of rules: *strict rules*, *defeasible rules*, and *defeaters*. Rules are also of two *kinds*:

- *Constitutive rules* (non-deontic, counts-as rules) R^{C} model constitutive statements;
- *Deontic rules* model prescriptive behaviours, which are either *obligation rules* R^{O} determining when and which obligations are in force, or *permission rules* representing *strong* (or *explicit*) permissions R^{P}.

Lastly, the *superiority* relation, $> \subseteq R \times R$, solves conflicts among rules' conclusions.

Following the ideas of [8], obligation rules gain more expressiveness with the *compensation operator* $\otimes$ for obligation rules, which is to model reparative chains of obligations. Intuitively, $a \otimes b \otimes c$ means that a is the primary obligation, but if, for some reason, we fail to comply with a, then b becomes the new obligation in force, and so on for c if we also fail with b. This operator, called $\otimes$-expressions, is hence used to build chains of "preference reparations" (c is still acceptable but less preferred than b).

Definition 2 (Rule)**.** *A* rule *is an expression of the form* $\alpha\colon A(\alpha) \hookrightarrow_{\Box} C(\alpha)$*, where*

1. $\alpha \in$ Lab *is the unique name of the rule;*
2. $A(\alpha) \subseteq$ Lit *is the set of antecedents;*
3. *An arrow* $\hookrightarrow \in \{\Rightarrow, \rightsquigarrow\}$ *denotes, respectively, defeasible rules, and defeaters;*
4. $\Box \in \{\mathsf{C}, \mathsf{O}, \mathsf{P}\}$*;*
5. $C(\alpha)$ *is the consequent, which is either*
   (a) a single plain literal $l \in$ PLit*, if (i)* $\hookrightarrow \equiv \rightsquigarrow$ *or (ii)* $\Box \in \{\mathsf{C}, \mathsf{P}\}$*, or*
   (b) an $\otimes$*-expression, if* $\Box \equiv \mathsf{O}$*.*

If $\Box = \mathsf{C}$ then the rule is used to derive non-deontic literals (constitutive statements), whilst if $\Box$ is O or P then the rule is used to derive deontic conclusions (prescriptive statements). The conclusion $C(\alpha)$ is a single literal in case $\Box = \{\mathsf{C}, \mathsf{P}\}$, or an $\otimes$-expression when $\Box = \mathsf{O}$. Note that $\otimes$-expressions can only occur in prescriptive rule though we do not admit them on defeaters (see [5]).

We use some standard abbreviations on rule sets. The set of defeasible rules is $R_{\Rightarrow}$, the set of defeaters is R_{dft}. $R^{\Box}[l]$ is the set of rules with conclusion l and modality $\Box$, while $R^{\mathsf{O}}[l, i]$ denotes the set of obligation rules where l is the i-th element in the $\otimes$-expression. Given that the consequent of a rule is either a single literal or an $\otimes$-expression, in what follows, we are going to shorten the notation and use $l \in C(\alpha)$.

Definition 3 (Tagged modal formula)**.** *A* tagged modal formula *is an expression of the form* $\pm\partial_{\Box} l$*, with the following meanings*

- $+\partial_{\Box} l$*: l is* defeasibly provable *(short,* provable*) with mode* $\Box$*,*
- $-\partial_{\Box} l$*: l is* defeasibly refuted *(short,* refuted*) with mode* $\Box$*;*

Accordingly, the meaning of $+\partial_{\mathsf{O}} p$ is that p is provable as an obligation, and $-\partial_{\mathsf{P}} \neg p$ is that we have a refutation for the permission of $\neg p$. Similarly, for the other combinations.

Definition 4 (Proof)**.** *Given a defeasible deontic theory D, a* proof P of length m in D *is a finite sequence* $P(1), P(2), \ldots, P(m)$ *of tagged modal formulas, where the proof conditions defined in the rest of this paper hold.* $P(1..n)$ *denotes the first n steps of P.*

The notational convention '$D \vdash \pm\partial_{\Box} l$*' means that there is a proof P for* $\pm\partial_{\Box} l$ *in D.*

Core notions in DL are that of *applicability/discardability* of a rule. This paper uses the one developed in [9,6]. As knowledge in a defeasible theory is circumstantial, given a defeasible rule like '$\alpha\colon a, b \Rightarrow_{\Box} c$', there are four possible scenarios: the theory defeasibly proves both a and b, the theory proves neither, the theory proves one but not the other. Naturally, only in the first case, where both a and b are proved, we can use α to *support/try to conclude* $\Box c$.

Definition 5 (Applicability)**.** *Assume a defeasible deontic theory* $D = (F, R, >)$.

1. *Rule* $\alpha \in R^{\mathsf{C}} \cup R^{\mathsf{P}}$ *is* applicable *at* $P(n+1)$*, iff for all* $a \in A(\alpha)$
   (a) *if* $a \in$ PLit*, then* $+\partial_{\mathsf{C}} a \in P(1..n)$,
   (b) *if* $a = \Box q$*, then* $+\partial_{\Box} q \in P(1..n)$*, with* $\Box \in \{\mathsf{O}, \mathsf{P}\}$,
   (c) *if* $a = \neg\Box q$*, then* $-\partial_{\Box} q \in P(1..n)$*, with* $\Box \in \{\mathsf{O}, \mathsf{P}\}$.
2. *Rule* $\alpha \in R^{\mathsf{O}}$ *is* applicable at index i and $P(n+1)$ *iff Conditions 1a–1c hold, and*
   (d) $\forall c_j \in C(\alpha)$, $j < i$*, then* $+\partial_{\mathsf{O}} c_j \in P(1..n)$ *and* $+\partial_{\mathsf{C}} {\sim} c_j \in P(1..n)$.
3. *Rule* $\alpha \in R^{\mathsf{C}}$ *is* applicable *at* $P(n+1)$ *for* $+\partial_{\Box} l$ *where* $\Box \in \{\mathsf{O}, \mathsf{P}\}$*, iff*
   (e) $\alpha \in R^{\mathsf{C}}[l]$,
   (f) *for all* $a \in A(\alpha)$, $a \in$ PLit*, and* $A(\alpha) \neq \emptyset$,
   (g) $+\partial_{\Box} a \in P(1..n)$.

Note that *discardability* of a rule is obtained by applying the principle of *strong negation* to the definition of applicability [7], and thus omitted.

Condition 1 establishes that (a, b) every positive literal has been proved, and (c) every deontic negative literal has been rejected at a previous derivation step.

Condition 2 deals with $\otimes$-chains: a rule is applicable at a certain index when each element c_j before have been proved as obligation $+\partial_{\mathsf{C}} c_j$ and violated $+\partial_{\mathsf{O}} {\sim} c_j$.

Lastly, Condition 3 formalises *rule conversion* mentioned in Section 2, which is a way to derive a conclusion with a certain modality by using rules for another modality [9]. In our case, constitutive rules can be used to derive obligations and permissions. This is formalised by Condition 3, which reads very easily: there must be a constitutive rule whose none of the antecedents is an obligation or permission (hence all of them plain literals), if all such antecedents are proved as obligations (resp. permissions) then the rule becomes applicable in supporting its conclusion as an obligation (resp. permission). Therefore, if we change rule α of above as '$\alpha\colon a, \mathsf{O}b \Rightarrow_{\mathsf{C}} c$', then this new α cannot be used through conversation and may only be used to support 'a constitutive c'.

For space reasons, we provide conditions for $+\partial_{\mathsf{O}}$ and $+\partial_{\mathsf{P}}$ only, since (i) $-\partial_{\mathsf{O}}$ and $-\partial_{\mathsf{P}}$ can be obtained by applying the strong negation principle to the positive counterparts, and (ii) the proof conditions for constitutive statements are the standard for DL [2].

Definition 6 (Obligation Proof Conditions)**.**

$+\partial_{\mathsf{O}} l$*: If* $P(n+1) = +\partial_{\mathsf{O}} l$ *then*
(1) $\mathsf{O}l \in F$*, or*
(2) $\mathsf{O}{\sim}l, \neg\mathsf{O}l, \mathsf{P}{\sim}l, \neg\mathsf{P}l \notin F$*, and*
(3) $\exists \beta \in R^{\mathsf{O}}_{\Rightarrow}[l, i] \cup R^{\mathsf{C}}_{\Rightarrow}[l]$ *s.t.*
(3.1) β *is applicable at index* i *if* $\beta \in R^{\mathsf{O}}_{\Rightarrow}[l, i]$*, or*
β *is applicable for* $+\partial_{\mathsf{O}} l$ *if* $\beta \in R^{\mathsf{C}}_{\Rightarrow}[l]$*, and*
(3.2) $\forall \gamma \in R^{\mathsf{O}}[{\sim}l, j] \cup R^{\mathsf{P}}[{\sim}l] \cup R^{\mathsf{C}}[{\sim}l]$ *either*
(3.2.1) γ *is discarded at index* j *if* $\gamma \in R^{\mathsf{O}}[{\sim}l, j]$*, or*
(3.2.2) $\exists \zeta \in R^{\mathsf{O}}[l, k] \cup R^{\mathsf{C}}_{\Rightarrow}[l]$ *s.t.*
(3.2.2.1) ζ *is applicable at index* k *if* $\zeta \in R^{\mathsf{O}}_{\Rightarrow}[l, k]$ *or*
ζ *is applicable for* $+\partial_{\mathsf{O}} l$ *if* $\zeta \in R^{\mathsf{C}}_{\Rightarrow}[l]$*, and*
(3.2.2.2) $\zeta > \gamma$.

Definition 7 (Permission Proof Conditions)**.**

$+\partial_{\mathsf{P}} l$: *If* $P(n+1) = +\partial_{\mathsf{P}} l$ *then*
(1) $\mathsf{P}l \in F$, *or*
(2) $\neg\mathsf{P}l, \mathsf{O}{\sim}l \notin F$, *and*
(3) $+\partial_{\mathsf{O}} l \in P(1..n)$, *or*
(4) $\exists \beta \in R^{\mathsf{P}}_{\Rightarrow}[l] \cup R^{\mathsf{C}}_{\Rightarrow}[l]$ *s.t.*
(4.1) β *is applicable if* $\beta \in R^{\mathsf{P}}_{\Rightarrow}[l]$ *or*
β *is applicable for* $+\partial_{\mathsf{P}} l$ *if* $\beta \in R^{\mathsf{C}}_{\Rightarrow}[l]$, *and*
(4.2) $\forall \gamma \in R^{\mathsf{O}}[{\sim}l, j] \cup R^{\mathsf{C}}_{\Rightarrow}[{\sim}l]$ *either*
(4.2.1) γ *is discarded at index* j *if* $\gamma \in R^{\mathsf{O}}[{\sim}l, j]$, *or*
(4.2.2) $\exists \zeta \in R^{\mathsf{P}}[l] \cup R^{\mathsf{O}}[l, k] \cup R^{\mathsf{C}}_{\Rightarrow}[l]$ *s.t.*
(4.2.2.1) ζ *is applicable at index* k *if* $\zeta \in R^{\mathsf{O}}[l, k]$,
or for $+\partial_{\mathsf{P}} l$ *if* $\zeta \in R^{\mathsf{C}}_{\Rightarrow}[l]$ *and*
(4.2.2.2) $\zeta > \beta$.

The set of positive and negative conclusions of a theory is called *extension*. The extension of a theory is computed based on the literals that appear in it; more precisely, the literals in the Herbrand Base of the theory $HB(D) = \{l, {\sim}l \in \text{PLit} \mid l \text{ appears in } D\}$.

Definition 8 (Extension)**.** *The* extension $E(D)$ *of a defeasible deontic theory* D *is*

$$E(D) = (+\partial_{\mathsf{C}}, -\partial_{\mathsf{C}}, +\partial_{\mathsf{O}}, -\partial_{\mathsf{O}}, +\partial_{\mathsf{P}}, -\partial_{\mathsf{P}}),$$

where $\pm\partial_{\Box} = \{l \in HB(D) \mid D \vdash \pm\partial_{\Box} l\}$, *with* $\Box \in \{\mathsf{C}, \mathsf{O}, \mathsf{P}\}$.

Theorem 1. *[See [5,9]] Given a defeasible theory* D, *its extension* $E(D)$ *can be computed in time polynomial to the size of the theory.*

4. Normative Explanation: The Formal Definition

As outlined in the previous sections, we explore the idea of stable deontic explanation by identifying those facts that ensure to prove a certain deontic conclusion. More precisely,

- Facts are added to an initial theory D_{init} and used to explain deontic conclusions in the resulting theory;
- We impose that only literals that do not appear as a consequence of any rule can be admissible facts for our purpose (factual literals).

The output theory (obtained by adding factual literals), as well as the whole operations, must thus satisfy certain properties.

Definition 9 (Admissible factual literals)**.** *Given (an initial) theory* $D_{init} = (\emptyset, R, >)$, *we define the set of* admissible factual literals *(shortly,* factual literals*) as*

$$\{a, \neg a, \mathsf{O}b, \mathsf{O}\neg b, \mathsf{P}c, \mathsf{P}\neg c \mid R^{\mathsf{C}}[a] \cup R^{\mathsf{C}}[\neg a] = \emptyset,$$
$$R^{\mathsf{O}}[b, i] \cup R^{\mathsf{C}}[b] \cup R^{\mathsf{O}}[\neg b, i] \cup R^{\mathsf{C}}[\neg b] = \emptyset$$
$$R^{\mathsf{O}}[c, i] \cup R^{\mathsf{P}}[c] \cup R^{\mathsf{C}}[c] \cup R^{\mathsf{O}}[\neg c, i] \cup R^{\mathsf{P}}[\neg c] \cup R^{\mathsf{C}}[\neg c] = \emptyset\}.$$

We say that an admissible factual literal l *is* deontic *iff* $l \in \text{ModLit}$.

It follows that the set of factual literals is the set of literals for which there are no rules; consequently, such literals can only be derived if they are facts of the theory.

Definition 10 (Consistent set of literals)**.** *A set of literals is* consistent *if it does not contain any pair of literals* $(p, \neg p)$, $(\Box p, \Box\neg p)$, $(\Box p, \neg\Box p)$, $(\mathsf{O}p, \mathsf{P}{\sim}p)$, *or* $(\mathsf{O}p, \neg\mathsf{P}p)$.

Example 2. *Assume D_{init} is $(\emptyset, R, \emptyset)$, with*

$$R = \{\alpha\colon a \Rightarrow_{\mathsf{C}} z, \zeta\colon z \Rightarrow_{\mathsf{O}} l, \beta\colon \mathsf{O}b \Rightarrow_{\mathsf{P}} \neg z, \gamma\colon \neg\mathsf{P}z \Rightarrow_{\mathsf{O}} l \otimes c, \delta\colon d, \mathsf{P}z \Rightarrow_{\mathsf{O}} \neg c\}.$$

Here, literals such as a, d, $\mathsf{O}b$, l, and $\neg z$ are (admissible) factual literals, whilst z, $\mathsf{O}l$, $\mathsf{P}\neg z$, $\mathsf{O}\neg c$, and $\mathsf{O}c$ are not.

Secondly, the output theory must be stable, i.e., consistently adding facts does not change the provability of the target literal. To formalise when a theory is stable, we firstly define which characteristics the output theory must satisfy, and we name such "valid" output theories *normative cases*.

Definition 11 (Normative Case)**.** *Let theory $D_{init} = (\emptyset, R, >)$ be the initial theory and $l \in \text{Lit} \cup \text{ModLit}$ be the target literal, we say that a theory $D = (F, R, >)$ is a* normative case for l of D_{init} *iff*

1. *F is consistent,*
2. *For all $f \in F$, f is a factual literal, and*
3. *$D \vdash +\partial_{\Box} p$ if $l = \Box p$ with $\Box \in \{\mathsf{O}, \mathsf{P}\}$, or $D \vdash +\partial_{\mathsf{C}} p$ if $l = p$.*

Definition 12 (Stable Normative Case)**.** *Given the initial theory $D_{init} = (\emptyset, R, >)$ and the target l, we say that a theory $D = (F, R, >)$ is*

1. *A* stable normative case for l of D_{init} *iff (1) D is a case for l (of D_{init}), and (2) for all $D' = (F', R, >)$ s.t. if $F \subsetneq F'$ and F' is consistent, then $D' \vdash +\partial_{\Box} p$ if $l = \Box p$, or $D' \vdash +\partial_{\mathsf{C}} p$ if $l = p$;*
2. *A* deontically stable normative case for l of D_{init} *iff (1) D is a case for l (of D_{init}), and (2) for all $D' = (F', R, >)$ s.t. if $F \subsetneq F'$, $F' \backslash F \subseteq \text{ModLit}$, and F' is consistent, then $D' \vdash +\partial_{\Box} p$ if $l = \Box p$, or $D' \vdash +\partial_{\mathsf{C}} p$ if $l = p$.*

We thus observe that, in contrast with the non-normative case [4], two sub-types of normative cases can be distinguished:

- A general case where a certain conclusion follows whatever facts are added to the initial theory;
- A deontic case where a certain conclusion follows whatever deontic facts (such as $\mathsf{O}p$ or $\neg\mathsf{P}q$) are added to the initial theory, but it is not ensured the stability if non-deontic facts are added.

Example 3. *Consider the theory D_{init} as in Example 2. The case theory $D = (F = \{\mathsf{P}a\}, R, \emptyset)$ is not stable for $\mathsf{P}z$ as $D' = (F' = \{\mathsf{P}a, \mathsf{O}b\}, R, \emptyset)$ proves $-\partial_{\mathsf{P}} z$. On the contrary, theories where the set of facts is $\{\mathsf{P}a, \mathsf{O}b, \sim l\}$ are stable normative cases for $\mathsf{O}c$.*

Suppose we add the rule '$\epsilon\colon \neg\mathsf{P}z, e \Rightarrow_{\mathsf{P}} \sim c$'. Theories with set of facts $\{\mathsf{P}a, \mathsf{O}b, \sim l\}$ are deontically stable normative cases, but we can no longer hold that they are stable in general, since adding the non-deontic fact e would prevent deriving $\mathsf{O}c$.

Symmetric to the concept of case, the notion of *normative refutation case* is:

Definition 13 (Normative Refutation Case)**.** *Let $\Box \in \{\mathsf{O}, \mathsf{P}\}$. Given the initial theory $D_{init} = (\emptyset, R, >)$ and the target literal $l \in \text{Lit} \cup \text{ModLit}$, we say that a theory $D = (F, R, >)$ is a* normative refutation case for l of D_{init} *iff*

1. *F is consistent,*
2. *For all $f \in F$, f is a factual literal, and*

3. $D \vdash -\partial_{\Box} p$ *if* $l = \Box p$, *or* $D \vdash -\partial_{\mathsf{C}} p$ *if* $l = p$.

Definition 14 (Stable Normative Refutation Case)**.** *Given the initial theory* $D_{init} = (\emptyset, R, >)$ *and the target* l, *we say that a theory* $D = (F, R, >)$ *is*

1. *A* stable normative refutation case for l of D_{init} *iff (1)* D *is a normative refutation case for* l *(of* D_{init}*), and (2) for all* $D' = (F', R, >)$ *s.t. if* $F \subsetneq F'$ *and* F' *is consistent, then* $D' \vdash -\partial_{\Box} p$ *if* $l = \Box p$, *or* $D' \vdash -\partial_{\mathsf{C}} p$ *if* $l = p$;
2. *A* deontically stable normative refutation case for l of D_{init} *iff (1)* D *is a normative refutation case for* l *(of* D_{init}*), and (2) for all* $D' = (F', R, >)$ *s.t. if* $F \subsetneq F'$, $F' \backslash F \subseteq$ ModLit, *and* F' *is consistent, then* $D' \vdash -\partial_{\Box} p$ *if* $l = \Box p$, *or* $D' \vdash -\partial_{\mathsf{C}} p$ *if* $l = p$.

Clearly, the following result trivially holds.

Proposition 1. *For any theories* D_{init} *and* D, *if* D *is a stable (refutation) case for* l *of* D_{init}, *then* D *is a deontically stable (refutation) case for* l *of* D_{init}.

The notion of *unstable case* can be directly introduced, which is the situation when a case is not resilient to the addition of facts to the theory.

Definition 15 (Unstable Normative Case)**.** *Given the initial theory* $D_{init} = (\emptyset, R, >)$ *and the target* l, *we say that a theory* $D = (F, R, >)$ *is* (deontically) normative unstable for l of D_{init} *iff (1)* D *is a normative case for* l *(of* D_{init}*), and (2) there exists* $D' = (F', R, >)$ *s.t. if* $F \subsetneq F'$ *(if* $F \subsetneq F'$, $F' \backslash F \subseteq$ ModLit*) and* F' *is consistent, then* $D' \vdash +\partial_{\Box} {\sim} p$ *if* $l = \Box p$, *or* $D' \vdash +\partial_{\mathsf{C}} {\sim} p$ *if* $l = p$

Note that, naturally, D is "just" a case for l, and not a stable (refutation) case.

A final interesting property is to identify a stable normative explanation which optimises the degree of deontic compliance.

Definition 16 (Optimal Stable Explanation)**.** *Given a theory* D *and its extension*

$$E(D) = (+\partial_{\mathsf{C}}, -\partial_{\mathsf{C}}, +\partial_{\mathsf{O}}, -\partial_{\mathsf{O}}, +\partial_{\mathsf{P}}, -\partial_{\mathsf{P}}),$$

the degree of compliance $\mathbf{Degree}(D)$ of D *is* $|\mathbf{Compl}(D)|$ *where* $\mathbf{Compl}(D) = \{l | D \vdash +\partial_{\mathsf{O}} l$ *and* $D \vdash -\partial_{\mathsf{C}} {\sim} l\}$. *A theory* D *is an* optimal stable normative explanation *for a target literal* l *iff* D *is a stable normative case and there is no other stable normative case* D' *for* l *such that* $\mathbf{Degree}(D') \leq \mathbf{Degree}(D)$.

Example 4. *Consider again Example 2. A theory* D *with set of facts '*$F = \{\mathsf{P}a, \mathsf{O}b, {\sim} l\}$*' is an optimal stable normative cases for* $\mathsf{O}c$. *However, if we have in* R *also '*$\theta :\Rightarrow_{\mathsf{O}} c$*', such that* $\theta > \delta$, *then* D *is no longer optimal.*

5. Complexity Results

The problem of determining if a case is stable is intractable in standard propositional DL [4]. The same results hold for DDL and the proof already developed can be directly used here as well: it is enough to show that for each Defeasible Deontic Theory, an equivalent propositional Defeasible Theory can be defined. The transformation is based on the procedure given in [15] to reduce a Defeasible Deontic Theory into a conclusion equivalent Defeasible Theory.

Theorem 2. *[15, Theorem 40] There is a polynomial transformation from any theory in DDL into its counterpart in DL.*

This theorem allows us to extend the complexity results for stability in DL [4, Theorems 2, 3 and 4] to the case of DDL.

Theorem 3. *Given a Defeasible Theory and a case, the problem of determining if the case is stable is co-NP-complete.*

Theorem 4. *Given a Defeasible Theory and a refutation case, the problem of determining if the refutation case is stable is co-NP-complete.*

Theorem 5. *Given a Defeasible Theory and a case, the problem of determining if the case is unstable is NP-complete.*

6. Summary

We examined the notion of deontic stability. We used Defeasible Deontic Logic, a tractable computationally oriented logic for the formalisation of norms, to provide a formal definition of the stability problem. We proved that to determine if an extension of a case is stable is computationally intractable even when the underlying (legal) reasoning system is tractable. The result indicates that, in general, creating an automated question-answering system posing questions to a user to determine a legal status (e.g., to determine what set of facts warrants a given legal outcome) is not feasible without additional heuristics. Accordingly, having determined the complexity, we plan to investigate suitable heuristics and identify tractable instances of the stability problem.

References

[1] Peter Achinstein. *The Nature of Explanation*. Oxford University Press, Oxford, 1983.
[2] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. Representation results for defeasible logic. *ACM Trans. Comput. Log.*, 2(2):255–287, 2001.
[3] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387, 2020.
[4] Guido Governatori, Francesco Olivieri, Antonino Rotolo, and Matteo Cristani. Inference to the stable explanations. In *LPNMR 2022*, pages 245–258, Cham, 2022. Springer.
[5] Guido Governatori, Francesco Olivieri, Antonino Rotolo, and Simone Scannapieco. Computing strong and weak permissions in defeasible logic. *J. Philos. Log.*, 42(6):799–829, 2013.
[6] Guido Governatori, Francesco Olivieri, Simone Scannapieco, Antonino Rotolo, and Matteo Cristani. The rationale behind the concept of goal. *Theory Pract. Log. Program.*, 16(3):296–324, 2016.
[7] Guido Governatori, Vineet Padmanabhan, Antonino Rotolo, and Abdul Sattar. A defeasible logic for modelling policy-based intentions and motivational attitudes. *Log. J. IGPL*, 17(3):227–265, 2009.
[8] Guido Governatori and Antonino Rotolo. Logic of violations: A gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic*, 4:193–215, 2006.
[9] Guido Governatori and Antonino Rotolo. BIO logical agents: Norms, beliefs, intentions in defeasible logic. *Auton. Agents Multi Agent Syst.*, 17(1):36–69, 2008.
[10] Guido Governatori, Antonino Rotolo, and Giovanni Sartor. Logic and the law: Philosophical foundations, deontics, and defeasible reasoning. In *Handbook of Deontic Logic and Normative Systems*, volume 2, pages 657–764. College Publications, London, 2021.
[11] Masha Medvedeva, Michel Vols, and Martijn Wieling. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28:237–266, 2020.
[12] Tim Miller, Robert Hoffman, Ofra Amir, and Andreas Holzinger, editors. *Artificial Intelligence journal: Special issue on Explainable Artificial Intelligence (XAI)*, volume 307, 2022.
[13] Donald Nute. Defeasible reasoning. In *Proceedings of the Hawaii International Conference on System Science*, volume 3, pages 470–477, 1987.
[14] Joseph C. Pitt. *Theories of Explanation*. Oxford University Press, Oxford, 1988.
[15] Simone Scannapieco. *Towards a Methodology for Business Process Revision Under Norm and Outcome Compliance*. PhD thesis, Griffith University, Brisbane, Australia, 2014.