Deontic Defeasible Reasoning in Legal Interpretation

Two Options for Modelling Interpretive Arguments

Antonino Rotolo*
CIRSFID, University of Bologna
Italy

Guido Governatori†
NICTA
Australia

Giovanni Sartor
EUI and University of Bologna
Italy

ABSTRACT

This paper offers a new logical machinery for reasoning about interpretive canons. We identify some options for modelling reasoning about interpretations and show that interpretative argumentation has a distinctive structure where the claim that a legal text ought or may be interpreted in a certain way can be supported or attacked by arguments, whose conflicts may have to be assessed according to further arguments.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms

Theory

Keywords

Defeasible Logic, Legal Interpretation, Argumentation

1. INTRODUCTION

Legal doctrine and judicial practice distinguish among a number of canons for interpreting legal statutes, i.e., different rules that are employed in legal systems as patterns for constructing arguments aimed at justifying certain interpretations, while attacking other interpretations. [11], summarising the outcomes of a vast study on statutory interpretation, involving scholars from many different legal systems, distinguishes eleven types of arguments. A different list of interpretive arguments was developed by [15] and identifies fourteen types of arguments.

*Supported by the Unibo FARB 2012 project *Mortality Salience, Legal and Social Compliance, and Economic Behaviour: Theoretical Models and Experimental Methods.*
†NICTA is funded by the Australian Government and the Australian Research Council through the ICT Centre of Excellence program.

ICAIL'15 June 08 - 12, 2015, San Diego, CA, USA

http://dx.doi.org/10.1145/2746090.2746100.

This paper proposes a new logical machinery for modelling reasoning about interpretive canons and thus for justifying the choice of a certain canon and the resulting legal outcome over competing interpretations (see [12, 10]). [14] argued that an interpretive canon for statutory law can be expressed as follows: if provision n occurs in document D, n has a setting of S, and n would fit this setting of S by having interpretation a, then, n ought to be interpreted as a. For instance, the ordinary language canon has the following structure:[1] if provision n, stating that "Killing a man is punishable by no less than 21 years in prison", occurs in document D = Penal code, n has a setting of ordinary language, and n would fit this setting of ordinary language by having interpretation a = "Killing an adult male person is punishable by no less than 21 years in prison", then n ought to be interpreted as a.

In this paper we basically accept this research line and work according to the following intuitions.

INTUITION 1 (REASONING AND CANONS). *We analyse the logical structure of interpretive arguments (in the sense of [11]) using a rule-based logical system. In particular, interpretation canons are represented by defeasible rules, where*

- *antecedent conditions of interpretation rules can be of any type (assertions, obligations, etc.), including the fact that another canon is refuted or that another legal provision ought to be interpreted in a certain way;*
- *the conclusion of interpretation rules is an interpretive act leading to an interpretation of a certain provision n and thus to a sentence which expresses the result of such an interpretation and paraphrases n [6]. If n and n' are legal provisions, the following is an example of interpretation rule regarding n':*

  IF
  n ought to be interpreted literally as a, AND
  n is related with n', AND
  a entails a',
  THEN
  n' is interpreted by coherence as a'.

We will use these rules to devise a reasoning machinery that mirrors legal reasoning about interpretive canons. The resulting rule-based system is in line with the basic ideas inspiring the argumentation system by [12].

[1]This argument supports the option that a provision be interpreted according to the meaning a native speaker of a given language would ascribe to it.

Notice that the above intuition distinguishes the interpretive act from the result of the interpretation:

INTUITION 2 (A- AND O-INTERPRETATION). *We assume the distinction between interpretation as activity and as outcome [13, p. 117] (cf. [15, p. 39]):*

- *interpretation as activity (*A-interpretation*) (literal or from ordinary language, by coherence, etc.) views any argumentative canon as a means through which a certain meaning is ascribed to a legal provision, and*
- *interpretation as outcome (*O-interpretation*) is precisely the meaning obtained through a certain interpretive act and ascribed to the provision.*

The distinction between interpretation as activity and as outcome is well known in continental legal theory, and it was introduced precisely to capture cases where, e.g., one has legal reasons to prefer a certain interpretive canon over others even though all considered canons support the same interpretive outcome. In other words, an interpretive act $\mid$ *of* n *as a (A-interpretation of* n*) is a way to bring about that a (O-interpretation of* n*) is the case. For example, in Intuition 1, the A-interpretation of* n' *is the act interpreting* n' *by coherence, while the resulting O-interpretation from that act is* a'*, i.e., a sentence expressing the meaning attributed to* n' *through the interpretation by coherence.*

Since different competing canons can be employed, different conflicting rules can be accordingly applied for interpreting statutes. Interpretation rules are thus defeasible. As argued in [14], some priority criteria should be applied to interpretation rules [1]. Such criteria impose preference relations over conflicting interpretive acts and outcomes. In other words, to address interpretive conflicts, we need to assume that one of the conflicting arguments is stronger than its competitors. Some legal traditions provide indeed general criteria for addressing conflicts of arguments on the basis of their priorities: for instance, several continental legal systems explicitly state that literal interpretation ought to be preferred, or that an argument concerning constitutional values ought to prevail over a historical argument (e.g., an argument based on the intent of the historical legislator).

However, ranking among interpretive acts and canons can be applied also when such acts are not in conflict. Suppose, for example, that provision n can be interpreted as a by adopting an argument by analogy and one from substantive reasons;[2] if n is a provision of criminal law (but analogy is admissible whenever it favours the defendant), then the argument from substantive reasons ought to be preferred, even though both lead to read n as a.

INTUITION 3 (PREFERENCES OVER INTERPRETATIONS). *A standard priority relation [2] over interpretation rules can be introduced to handle and solve conflicts between different interpretation rules. Consider the following example:*

Rule1
IF
n *ought to be interpreted literally as* a, AND
n *is related with* n', AND
a *entails* a',
THEN
n' *is interpreted by coherence as* a'

Rule2
IF
n'' *ought to be interpreted literally as* $\neg a$, AND
n *is related with* n'', AND
$\neg a$ *entails* $\neg a'$,
THEN
n' *is interpreted by coherence as* $\neg a'$.

Here, we can handle the conflict by stating that **Rule1** $>$ **Rule2** *(or vice versa).*

Ranking among interpretive acts can be applied also when such acts are not in conflict. We will thus introduce an operator $\otimes$ *that can be used to make explicit in single rules this idea. For instance,*

IF
n *ought to be interpreted literally as* a, AND
n *is related with* n'
THEN
n' *is interpreted by coherence as* $a' \otimes$
$\otimes$ n' *is interpreted by analogy as* a'

means that the most preferred interpretation resulting in a is the one by coherence, but, if this is refuted, the second option is the interpretation by analogy. This does not require to only derive one interpretation resulting in a (other rules could first support interpretation by analogy of n*).*

Following some doctrinal and judicial practice, [14] argued that interpretive canons are defeasible rules licensing deontic interpretive claims, namely, the claim that a certain expression in a statute ought, ought not, may or may not be interpreted in a certain way. For example, art. 12 in the general provisions of the Italian civil code states that the literal interpretation of statues ought to be preferred and this option is nothing but an interpretive prescription. Here, we follow this intuition with some adjustments.

INTUITION 4 (OBLIGATORY INTERPRETATIONS). *An interpretation can be* admissible *or* obligatory*. In the case of A-interpretations, for instance, an interpretive act* $\mid$ *of* n *(A-interpretation of* n*) is admissible, if it is provable using a defeasible interpretation rule; it is obligatory, if this interpretation of* n *is the only one admissible. Similarly for O-interpretations. Indeed, consider the general provisions of the Italian civil code, which state at art. 12 that literal interpretation* $\mid_{lit}$ *ought to be preferred: this would support that such interpretation is obligatory, unless another interpretation prevails. We have two options here:*

- *other conflicting interpretations can be derived, thus requiring to check if literal interpretation overrides the other options; if it does not, then the interpretation at stake is not even admissible;*
- *other non-conflicting interpretations can be provable; if they are, the interpretation at stake is only admissible, otherwise, it is obligatory.*

[2] An argument from substantive reasons states that, if there is some goal that can be considered to be fundamentally important to the legal system, and if the goal can be promoted by one rather than another interpretation of the statutory provision, then the provision should be interpreted in accord with the goal.

On the basis of the above intuitions, we will offer two options for modelling reasoning about interpretations: a defeasible logic for reasoning about the interpretation of abstract, non-analysed provisions and of structured provisions.

INTUITION 5 (ABSTRACT OR STRUCTURED PROVISIONS). *A provision n is* abstract *if it is taken in its sentential entirety for interpretive purposes, i.e., as a non-analysed sentence without considering its internal (logical) structure. In our paper,* Option 1 *amounts to interpreting n by ascribing to n, intended as an abstract provision, a sentential meaning that can be expressed by another sentence paraphrasing this provision as a whole.*

Rather, a provision n' is logically structured *if it corresponds to a linguistic sentence having the structure of a rule $a_1, \ldots, a_n \Rightarrow b$: this means that n' is semi-interpreted provision, since expressing the logical structure of n' requires an interpretive effort on the original textual version of n'. In our paper,* Option 2 *amounts to interpreting n' by considering the components $a_1, \ldots, a_n, b$ of n' and ascribing to them a meaning as already explained above.*

The above intuitions are implemented adjusting the framework in [7], which is a Modal Defeasible Logic [8] extended with the operator $\otimes$. The logic is a significant extension of standard Defeasible Logic [2], which however preserves linear computational complexity (like standard Defeasible Logic).

The layout of the paper is as follows. Section 2 offers a Modal Defeasible Logic for Option 1 (see Intuition 5 above), i.e., a machinery for reasoning about the interpretation of abstract provisions: Section 2.1 presents the formal language, Section 2.2 introduces some preliminary concepts, while Sections 2.3 and 2.4 offer a complete proof theory for A-interpretations and O-interpretations (see Intuition 2). Section 3 similarly offers a Modal Defeasible Logic for reasoning about the interpretation of structured provisions. Some discussion end the paper.

2. OPTION 1: ABSTRACT NORMS

Let us use Defeasible Logic to reason about interpretive arguments that handle the overall meaning of legal provisions intended as argumentative, abstract (i.e., non-analysed) logical units. We thus have the following basic components (among others):

- a set of legal provisions $n_1, n_2, \ldots$ to be interpreted;
- as set of literals $a, b, \ldots$, corresponding to any sentences, which can be used to offer a sentential meaning to any provision n (a literal a is the meaning of provision n);
- a set of interpretative acts or interpretations $\mathsf{I}_1, \mathsf{I}_2, \ldots$ (literal interpretation, teleological interpretation, etc.) that return for any legal provision a sentential meaning for it;
- a set of rules for reasoning about ordinary meanings, i.e., literals, and rules encoding interpretive arguments (i.e., rules that state what interpretive act can be obtained under suitable conditions).

2.1 Formal Language

The language of our logic is formally defined as follows.

DEFINITION 1 (LANGUAGE, OPTION 1). *Let* PROP $= \{a, b, \ldots\}$ *be a set of propositional atoms,* NORM $= \{n_1, n_2, \ldots\}$ *a set of legal provisions,* $\mathsf{INTR} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$ *a set of interpretation functions (for example, denoting literal interpretation, etc.),* MOD $= \{\mathsf{Obl}, \mathsf{Adm}\}$ *a set of modal operators where* Obl *is the modality for denoting obligatory interpretations and interpretation outcomes and* Adm *for denoting the admissible ones, and $\sqsubseteq$ and $\approx$ the relations denoting semantic inclusion and equivalence.*[3]

1. *The set* Lit $=$ PROP $\cup \{\neg p \,|\, p \in$ PROP$\}$ *denotes the set of* literals.
2. *The* complementary of a literal q *is denoted by $\sim q$; if q is a positive literal p, then $\sim q$ is $\neg p$, and if q is a negative literal $\neg p$, then $\sim q$ is p.*
3. *The set* ModLit $= \{\Box a, \neg\Box a | a \in$ Lit$, \Box \in$ MOD$\}$ *denotes the set of modal literals.*
4. *The set* Int $= \{\mathsf{I}_i(n, a), \neg\mathsf{I}_i(n, a) | \exists \mathcal{I}_i : \text{NORM} \mapsto \text{Lit} \in \mathsf{INTR} : \mathcal{I}_i(n) = a\}$ *denotes the set of* interpretive acts *and their negations: an expression $\mathsf{I}_i(n, a)$, for instance, means that the interpretation I_i of provision n returns that the literal a is the case.*
5. *The* complementary of an interpretation ϕ *is denoted by $\sim\phi$ and is defined as follows:*

ϕ	$\sim\phi$
$\mathsf{I}_i(n, a)$	$\sim\mathsf{I}_i(n, a) \in \{\neg\mathsf{I}_i(n, a), \mathsf{I}_i(n, b), \mathsf{I}_j(n, c) \mid a \neq b$, *and either* $a \not\sqsubseteq c$ *or* $c \not\sqsubseteq a\}$
$\neg\mathsf{I}_i(n, a)$	$\sim\neg\mathsf{I}_i(n, a) \in \{\mathsf{I}_i(n, a), \mathsf{I}_i(n, c) \mid a \approx c\}$

   We will also use the notation $\pm\mathsf{I}_i(n, a)$ to mean respectively $\mathsf{I}_i(n, a)$ and $\sim\mathsf{I}_i(n, a)$. Hence, $\sim \pm \mathsf{I}_i(n, a)$ means $\mp\mathsf{I}_i(n, a)$.
6. *The set of* qualified interpretations *is* ModIntr $= \{\Box\phi, \neg\Box\phi \,|\, \phi \in$ Int$, \Box \in$ MOD$\}$.
7. *The* complementary of a modal literal or qualified interpretation l *is defined as follows ($\phi \in$ Lit $\cup$ Int):*

$\mathbf{l}$	$\sim\mathbf{l}$
$\mathsf{Obl}\phi$	$\sim\mathsf{Obl}\phi \in \{\neg\mathsf{Obl}\phi, \mathsf{Obl}\sim\phi, \mathsf{Adm}\sim\phi, \neg\mathsf{Adm}\phi\}$
$\neg\mathsf{Obl}\phi$	$\sim\neg\mathsf{Obl}\phi = \mathsf{Obl}\phi$
$\mathsf{Adm}\phi$	$\sim\mathsf{Adm}\phi \in \{\neg\mathsf{Adm}\phi, \mathsf{Obl}\sim\phi\}$
$\neg\mathsf{Adm}\phi$	$\sim\neg\mathsf{Adm}\phi = \mathsf{Adm}\phi$

REMARK 1. *A complementary of an interpretation I_i is another interpretation that is* necessarily incompatible *with I_i. Given any $\mathsf{I}_i(n, a)$ ("n is interpreted by I_i as a"), its propositional negation $\neg\mathsf{I}_i(n, a)$ ("it is false that n is interpreted by I_i as a") leads to an incompatibility. But other cases should be considered. First of all, the same interpretation (literal, teleological, etc.) of the same provision cannot give two (or*

[3] A simple characterisation of $\sqsubseteq$ and $\approx$ is given in Definition 5 As we will see, if a rule says $a \to b$, then $a \sqsubseteq b$, but the other way around does not necessarily hold. Similarly, if $a = b$, then $a \equiv b$, while the opposite direction is not necessarily the case.

more) distinct results, even if these ones are somehow compatible. For instance, let us consider the following provision from the Italian penal code:

> *Art. 575. Homicide. Whoever causes the death of a* man *[*uomo*] is punishable by no less than 21 years in prison.*

The literal interpretation (denoted, e.g., by I_1*) of art. 575 cannot lead at the same time to reading "man", e.g., as "adult male human being" and as "human being", even though being an adult male human being entails being a human being. If we consider two distinct interpretations* I_i *and* I_j *of the same provision, things are slightly different because we resort to two different interpretive arguments. Indeed, suppose to fictionally change art. 575 and replace "man" with "smart Italian person": if the lawmaker likes people from Florence, one may resort to a psychological argument [14, 15] and support the interpretation* I_i *driven by the actual intent of punishing those who kill Florentine people. On the other side, if a naturalistic argument [14, 15] supporting an interpretation* I_j *leads to reading "smart Italian person" as "person living in a nice place", the fact that* I_i *and* I_j *are necessarily incompatible depends, for instance, on whether being Florentine does not entail to live in a nice place: if this is the case those interpretations cannot be compatible. This analysis is partial and debatable: more logical relations between the propositional content of provisions can be identified [3, 4], thus allowing to refine the concept of conflicting interpretation. Also, one may argue that any interpretations* $\mathsf{I}_i(n, a)$ *and* $\mathsf{I}_j(n, a')$ *of* n *are in conflict [14] whenever* $a \neq a'$*, irrespective of any possible logical relation (entailment, semantic overlapping, etc.) between* a *and* a'*. In fact, while we will rely on the above analysis to illustrate the logical machinery, we do not commit to it: different concepts of complementary interpretation can be introduced. More comments will be given in the conclusions of the paper.*

We introduce a preference operator $\otimes$ over interpretations. Indeed, when more interpretations of a provision are available, one may wonder what is the most preferred one [14]. For example, an expression like $\mathsf{I}_i(n, a) \otimes \mathsf{I}_j(n, a) \otimes \mathsf{I}_k(n, b)$ means that the most preferred interpretation of provision n is a via I_i; if this is not admissible, then the second best choice is the same result a via a different interpretation type I_j; if this last is not admissible, too, n is interpreted as b using interpretation I_k. This operator is thus used to build chains of preferences, called $\otimes$-expressions. The formation rules for well-formed $\otimes$-expressions are:

1. $\forall n \in$ NORM, $\forall \mathsf{I}_i \in$ Int, and $\forall a \in$ Lit, each $\mathsf{I}_i(n, a)$ is an $\otimes$-expression;
2. If $\phi_1, \ldots, \phi_n \in$ Int, then $\phi_1 \otimes \cdots \otimes \phi_n$ is an $\otimes$-expression;
3. Nothing else is an $\otimes$-expression.

Example 1. *Let us consider again art. 575 of the Italian penal code. The original version speaks of "the death of a* man*", i.e., in the ordinary sense, "an adult male human being".*[4] *It goes without saying that this provision does not confine to adult male persons only. However, for the sake of argument let us assume that the literal interpretation of art. 575 (denoted by* I_1*) corresponds to accepting the ordinary meaning of the term "man [*uomo*]"; then, we would have that* $\mathsf{I}_1(art.575, a)$*, where* $a =$*"Whoever causes the death of* a adult male *person is punishable by no less than 21 years in prison". Other options would rather interpret art. 575 by covering any person (not only adult men): this could be the same reasonable conclusion of arguments such as those from general principles (*$\mathsf{I}_2(art.575, b)$*) and from substantive reasons (*$\mathsf{I}_3(art.575, b)$*),where* $b =$*"Whoever causes the death of a* person *is punishable by no less than 21 years in prison". Suppose* I_2 *is the best choice,* I_3 *the second best choice, and* $\neg\mathsf{I}_1$ *the least preferred option; the following* $\otimes$*-expression represents this preference ordering:*

$$\mathsf{I}_2(art.575, b) \otimes \mathsf{I}_3(art.575, b) \otimes \neg\mathsf{I}_1(art.575, a).$$

We stipulate that $\otimes$ obeys the following properties:

1. $\phi \otimes (\psi \otimes \pi) = (\phi \otimes \psi) \otimes \pi$ (associativity);
2. $\bigotimes_{i=1}^{n} \phi_i = (\bigotimes_{i=1}^{k-1} \phi_i) \otimes (\bigotimes_{i=k+1}^{n} \phi_i)$ where there exists j such that $\phi_j = \phi_k$ and $j < k$ (duplication and contraction on the right).

Given an $\otimes$-expression A, the *length* of A is the number of interpretations in it. Given an $\otimes$-expression $A \otimes \phi \otimes C$ (where A and C can be empty), the *index* of ϕ is the length of $A \otimes \phi$. We also say that ϕ appears at index n in $A \otimes \phi$ if the length of $A \otimes \phi$ is n.

We adopt the standard definitions in Defeasible Logic of *strict rules*, *defeasible rules*, and *defeaters*[5] [2] with the following modifications:

strict rules, encoding the monotonic part of the knowledge base, are used to reason about basic linguistic meanings in ordinary parlance and thus contain only literals in their antecedents and consequents; we denote them as usual with the arrow $\rightarrow$ and mark them with a superscript M;[6]

the non-monotonic part of the knowledge base (defeasible rules and defeaters) is used to reason about the interpretations of provisions; these rules contain literals, interpretations and qualified interpretations in their antecedent, and $\otimes$-expressions in their consequents; we denote them as usual with the arrows $\Rightarrow$ and $\rightsquigarrow$ and mark them with a superscript I .

Definition 2 (Rules, Option 1). *Let* Lab *be a set of arbitrary labels. Every rule is of the type*

$$r : A(r) \hookrightarrow^{X} C(r)$$

where

1. $r \in$ Lab *is the name of the rule, and either*
2. *if* r *is a* strict rule*, then* r *is a* meaning rule *such that*
   - $A(r) = \{a_1, \ldots, a_n\}$*, the* antecedent *(or* body*) of the rule, is the set of the premises of the rule (alternatively, it can be understood as the conjunction of all the literals in it). Each* a_i *is a literal;*

[4] Entry "Man", *Online Cambridge English Dictionary*, 2014.

[5] A defeater is a rule which prevents opposite conclusions without allowing to positively deriving anything.

[6] One may notice that any suitable logic modelling the meanings of ordinary language should be non-monotonic: a sentence like "all birds are flying animals" admits exceptions. We use only strict rules to simplify proof theory.

- $\hookrightarrow = \rightarrow^M$;
- $C(r)$ *is the* consequent *(or* head*) of the rule, which is a literal;*

3. *otherwise,* r *is an* interpretation rule *such that*

4. $A(r) = \{\phi_1, \ldots, \phi_n\}$*, the* antecedent *(or* body*) of the rule is such that each* ϕ_i *is either a literal* $l \in$ Lit*, a modal literal* $Y \in$ ModLit*, or a qualified interpretation* $X \in$ ModIntr*;*

5. $\hookrightarrow \in \{\Rightarrow^I, \rightsquigarrow^I\}$ *denotes the type of the rule. If* $\hookrightarrow$ *is* $\Rightarrow^I$*, the rule is a* defeasible rule*, while if* $\hookrightarrow$ *is* $\rightsquigarrow^I$*, the rule is a* defeater*;*

6. $C(r)$ *is the* consequent *(or* head*) of the rule, which is an* $\otimes$*-expression; if* $\hookrightarrow$ *is* $\rightsquigarrow$*, then* $C(r)$ *is a single interpretation, i.e., an* $\otimes$*-expression of length* 1*.*

Example 2. *Let us consider the following examples of meaning (strict) rules:*

$$\begin{array}{ll} r_1: & kill\_man \rightarrow^M kill\_person \\ r_2: & kill\_male, kill\_adult \rightarrow^M kill\_man \\ r_3 & kill\_female \rightarrow^M \neg kill\_male \\ r_4 & kill\_newborn \rightarrow^M \neg kill\_adult \end{array}$$

Consider now that paragraph 1 of art. 3 of the Italian constitution reads as follows:

> *All citizens have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*

The most preferred interpretation I_3 *(interpretation from substantive reasons) of it leads to c, which corresponds to the following sentence:*

> *All persons have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*

The following interpretation defeasible rule could be in order:

$$r_5: \quad kill\_adult, kill\_female, \mathsf{Obl}\,\mathsf{I}_3(art.3, c) \Rightarrow^I \mathsf{I}_2(art.575, b) \otimes \mathsf{I}_3(art.575, b) \otimes \neg \mathsf{I}_1(art.575, a).$$

In other words, if art. 3 of the Italian constitution states formal equality before the law without regard also to gender identity, then b is the best interpretation outcome of art. 575 of the penal code, with I_2 *(e.g., interpretation from general principles) as preferred over* I_3 *(say, interpretation from substantive reasons). In the light of art. 3, if these two interpretive options are refuted by any stronger interpretive conclusions, then the last sub-ideal option is to reject the interpretation* I_1 *from ordinary meaning (according to which only the homicide of adult male persons is punishable!).*

Given a set of rules R, we use the following abbreviations for specific subsets of rules:

- R^I_{deft} denotes the set of all defeaters in the set R (since defeaters can only be of type I, we will omit the superscript in the remainder);
- $R^I[\phi, n]$ is the set of rules where the interpretation ϕ appears at index n in the consequent. The set of (defeasible) rules where ϕ appears at any index n is denoted by $R^I[\phi]$ (since rules with $\otimes$-expressions can only be of type I, we will omit the superscript in the remainder);
- $R^M[l]$ denotes the set of all strict rules with head l.

Definition 3 (Interpretation theory). *An* Interpretation Theory *is a structure* $D = (F, R, >)$*, where* F*, the set of facts, is a set of literals, modal literals, and qualified interpretations,* R *is a set of rules and* $>$*, the superiority relation, is a binary relation over* R*.*

An interpretation theory corresponds to a knowledge base providing us with interpretive arguments about legal provisions. The superiority relation is used for conflicting rules, i.e., rules whose conclusions are complementary. We do not impose any restriction on the superiority relation: it just determines the relative strength of two rules.

Example 3. *Consider art. 575 of the Italian penal code, art. 3 of the Italian constitution. Also consider art. 578 (of the penal code):*

> *Art. 578. Infanticide. The mother who causes the death of her newborn baby immediately after birth [. . .], when this act is committed in conditions of material or moral distress, is punishable with a sentence between 4 and 12 years of prison.*

Assume that

$$\begin{array}{lcl} a & = & \textit{Whoever causes the death of a adult male person is punishable by no less than 21 years in prison} \\ a' & = & \textit{Whoever causes the death of a male person is punishable by no less than 21 years in prison} \\ b & = & \textit{Whoever causes the death of a person is punishable by no less than 21 years in prison} \\ c & = & \textit{All persons have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.} \end{array}$$

$$\begin{array}{lcl} \mathsf{I}_1 & = & \textit{Literal interpretation or from ordinary meaning} \\ \mathsf{I}_2 & = & \textit{Interpretation from general principles} \\ \mathsf{I}_3 & = & \textit{Interpretation from substantive reasons} \\ \mathsf{I}_4 & = & \textit{Interpretation by coherence} \end{array}$$

The following theory reconstructs an interpretive toy scenario in the Italian legal system.

$$\begin{array}{l} F = \{kill\_adult, kill\_female, \mathsf{Obl}\,\mathsf{I}_3(art.3, c), \mathsf{Obl}\,\mathsf{I}_1(art.578, d)\} \\ R = \{r_5 : kill\_adult, kill\_female, \mathsf{Obl}\,\mathsf{I}_3(art.3, c) \Rightarrow^I \\ \qquad \mathsf{I}_2(art.575, b) \otimes \mathsf{I}_3(art.575, b) \otimes \neg \mathsf{I}_1(art.575, a), \\ \qquad r_6 : \mathsf{Obl}\,\mathsf{I}_1(art.578, d) \Rightarrow^I \mathsf{I}_1(art.575, a) \otimes \mathsf{I}_4(art.575, a), \\ \qquad r_7 :\Rightarrow^I \mathsf{I}_1(art.575, a'),\ \ r_8 : b \rightarrow^M a\} \\ > = \{r_6 > r_5\} \end{array}$$

Rule r_5 *has been already introduced above. Rule* r_6 *states that, in case one kills an adult person and art. 578 ought to be interpreted literally, then we have a reason to interpret art. 575 by coherence (with respect to art. 578) as a. Rule*

r_7 establishes by default that art. 575 be literally interpreted as a'. Finally, rule r_8 is a meaning rule saying that b entails a, i.e., that a provision punishing whoever causes the death of a person entails that a provision should be in the system that punishes whoever causes the death of a an adult male person.

2.2 Proof Theory: Preliminaries

This section presents some preliminary concepts for reasoning about interpretations: the notions of proof, semantic inclusion, and rule applicability.

DEFINITION 4. *[Proofs] A* proof *P in an interpretation theory D is a linear sequence $P(1)\dots P(n)$ of* tagged expressions *in the form of $+\Delta^M q$, $-\Delta^M q$ (with $q \in$ Lit), $+\partial^I_\Box \phi$ and $-\partial^I_\Box \phi$ (with $\phi \in$ Int and $\Box \in$ MOD), $+\partial^\Box l$ and $-\partial^\Box l$ (with $l \in$ Lit and $\Box \in$ MOD), where $P(1)\dots P(n)$ satisfy the proof conditions given in Definitions 9–16.*

The tagged literal $+\Delta^M q$ means that q is strictly or monotonically provable in D using only facts and strict rules, while $-\Delta^M q$ means that there is no strict of monotonic proof for q in D. The tagged interpretation $+\partial^I_\Box \phi$ means that the interpretation ϕ is *defeasibly provable* in D with modality $\Box$, while $-\partial^I_\Box \phi$ means that ϕ is *defeasibly refuted* with modality $\Box$. The tagged literal $+\partial^\Box l$ means that l is *defeasibly provable* in D with modality $\Box$,[7] while $-\partial^\Box l$ means that l is *defeasibly refuted* with modality $\Box$. The initial part of length n of a proof P is denoted by $P(1..n)$.

Proof theory for strict conclusions is trivially the same for Δ-conclusions in standard Defeasible Logic [2]. We omit it here. We only state the following, to define $\sqsubseteq$ and $\approx$:

DEFINITION 5 (SEMANTIC INCLUSION). *Let $D = (F, R, >)$ an interpretation theory. If $p, q \in$ Lit, we write $p \sqsubseteq_D q$[8] to say that the meaning of literal p is semantically included by q in D. If $\mathcal{P} = \{r | r \in R^M, p \in A(r)$ or $p \in C(r)\}$, $p \sqsubseteq_D q$ if*

- *$D \vdash +\Delta^M p$ and $D \vdash +\Delta^M q$, and*
- *$D_\sqsubseteq = (F_\sqsubseteq R^M_\sqsubseteq, >) \nvdash +\Delta^M q$ where (i) $F_\sqsubseteq = F \setminus \{p, q\}$ and (ii) $R^M_\sqsubseteq$ is such that $\{r | r \in R_\sqsubseteq\} \setminus \mathcal{P}$. We state that $p \approx_D q$ iff $p \sqsubseteq_D q$ and $q \sqsubseteq_D p$.*

In other words, a literal p is included by q in a theory D when both are obtained in D using meaning rules, but, if p and q are removed from the set of facts (if they are there) and all meaning rules having p in the head of body are removed as well, then q no longer follows from the theory. Hence, it means that p is decisive for having q. As we have already observed, semantic inclusion does not fully coincide with entailment via a single rule. Indeed, if $p \sqsubseteq q$, this could mean, for example, that $\{a_1, \dots, a_n, p\} \vdash_D q$, $\{a_1, \dots, a_n\} \nvdash_D q$, but also that $\{p\} \nvdash_D q$.

Let us work on the proof theory for deriving qualified interpretations and first consider the admissible ones (i.e., for any interpretation ϕ, those having the form $\mathsf{Adm}\phi$). The first thing to do is to define when a rule is applicable or discarded. A rule is *applicable* for an interpretation ϕ if ϕ occurs in the head of the rule, all literals in the antecedent are Δ^M-provable and all the qualified interpretations and modal literals in the antecedent have been defeasibly proved (with the appropriate modalities). On the other hand, a rule is *discarded* if at least one of the literals, modal literals, or of the qualified interpretations in the antecedent has not been proved. However, as interpretation ϕ might not appear as the first element in an $\otimes$-expression in the head of the rule, some additional conditions on the consequent of rules must be satisfied. Defining when a rule is applicable or discarded is essential to characterise the notion of provability for admissible interpretations ($\pm\partial^I_{\mathsf{Adm}}$).

DEFINITION 6. *A rule $r \in R^I$ is* body-applicable *in the proof P at $P(n+1)$ iff for all $a_i \in A(r)$:*

1. *if $a_i = \Box\psi$, $\psi \in$ Int, then $+\partial^I_\Box \psi \in P(1..n)$ with $\Box \in$ MOD;*
2. *if $a_i = \neg\Box\psi$ then $-\partial^I_\Box \psi \in P(1..n)$ with $\Box \in$ MOD;*
3. *if $a_i = \Box l$, $l \in$ Lit, then $+\partial^\Box l \in P(1..n)$;*
4. *if $a_i = \neg\Box l$, $l \in$ Lit, then $-\partial^\Box l \in P(1..n)$;*
5. *if $a_i = l \in$ Lit then $D \vdash +\Delta^M l$.*

A rule $r \in R^I$ is body-discarded *iff $\exists a_i \in A(r)$ such that*

1. *if $a_i = \Box\psi$, , $\psi \in$ Int, then $-\partial^I_\Box \psi \in P(1..n)$ with $\Box \in$ MOD;*
2. *if $a_i = \neg\Box\psi$, $\psi \in$ Int, then $+\partial^I_\Box \psi \in P(1..n)$ with $\Box \in$ MOD;*
3. *if $a_i = \Box l$, $l \in$ Lit, then $-\partial^\Box l \in P(1..n)$;*
4. *if $a_i = \neg\Box l$, $l \in$ Lit, then $+\partial^\Box l \in P(1..n)$;*
5. *if $a_i = l \in$ Lit then $D \vdash -\Delta^M l$.*

DEFINITION 7. *An interpretation rule $r \in R^I[\phi, j]$ such that $C(r) = \phi_1 \otimes \dots \otimes \phi_n$ is* applicable *for the interpretation ϕ at index j, with $1 \le j \le n$, in the condition for $\pm\partial_{\mathsf{Adm}}$ (in the proof P at $P(n+1)$) iff*

1. *r is body-applicable; and*
2. *for all $\phi_k \in C(r)$, $1 \le k < j$, $+\partial^I_{\mathsf{Adm}} \sim \phi_k \in P(1..n)$.*

Conditions (1) represents the requirements on the antecedent stated in Definition 6; condition (2) on the head of the rule states that each element ϕ_k prior to ϕ must be refuted as an admissible interpretation.

DEFINITION 8. *An interpretation rule $r \in R^I[\phi, j]$ such that $C(r) = \phi_1 \otimes \dots \otimes \phi_n$ is* discarded *for interpretation ϕ at index j, with $1 \le j \le n$ in the condition for $\pm\partial_{\mathsf{Adm}}$ iff*

1. *r is body–discarded; or*
2. *there exists $\phi_k \in C(r)$, $1 \le k < j$, such that $+\partial_{\mathsf{Adm}} \phi_k \in P(1..n)$.*

In this case, condition (2) states that there exists at least one explicit derived admissible interpretation prior to ϕ.

[7] As resulting here from a provable A-interpretation.
[8] When clear from the context, we omit the subscript.

2.3 Proof Theory: Interpretation Activities

We work in this subsection on the proof theory for reasoning about interpretation activities (A-Interpretations).

Let us define the proof conditions for $\pm\partial_{\mathsf{Adm}}$.

Definition 9 (Admissible A-Interpretations). *The proof condition of* defeasible provability for admissible A-interpretations *is*

$+\partial^I_{\mathsf{Adm}}$*: If* $P(n+1) = +\partial^I_{\mathsf{Adm}}\phi$ *then*
(1) $\mathsf{Adm}\phi \in F$ *or* $\mathsf{Obl}\phi \in F$*, or*
  (2.1) $\sim\mathsf{Adm}\phi \notin F$*, and*
  (2.2) $\exists r \in R[\phi, i]$ *such that* r *is applicable for* ϕ*, and*
  (2.3) $\forall s \in R[\sim\phi, j]$*, either*
    (2.3.1) s *is discarded for* $\sim\phi$*, or*
    (2.3.2) $\exists t \in R[\phi, k]$ *such that* t *is applicable for* ϕ *and* $t > s$.

To show that ϕ is defeasibly provable as an admissible interpretation, there are two ways: (1) $\mathsf{Adm}\phi$ or $\mathsf{Obl}\phi$ are a fact, or (2) $\mathsf{Adm}\phi$ must be derived by the rules of the theory. In the second case, three conditions must hold: (2.1) any complementary of $\mathsf{Adm}\phi$ does belong to the facts; (2.2) there must be a rule introducing the admissibility for ϕ which can apply; (2.3) every rule s for $\sim\phi$ is either discarded or defeated by a stronger rule for ϕ.

The strong negation [2] of Definition 9 gives the negative proof condition for admissible interpretations.

Definition 10 (Refuted A-Interpretations). *The proof condition of* defeasible refutability for admissible A-interpretations *is*

$-\partial^I_{\mathsf{Adm}}$*: If* $P(n+1) = -\partial^I_{\mathsf{Adm}}\phi$ *then*
(1) $\mathsf{Adm}\phi \notin F$ *and* $\mathsf{Obl}\phi \notin F$*, and*
  (2.1) $\sim\mathsf{Adm}\phi \in F$*, or*
  (2.2) $\forall r \in R[\phi, i]$ r *is discarded for* ϕ*, or*
  (2.3) $\exists s \in R[\sim\phi, j]$ *such that*
    (2.3.1) s *is applicable for* $\sim\phi$*, and*
    (2.3.2) $\forall t \in R[\phi, k]$*, either* t *is discarded for* ϕ *or* $t \not> s$.

Proof conditions for $\pm\partial_{\mathsf{Obl}}$ are much easier but we need to work on the fact that ϕ is an interpretation of any given provision n and we have to make explicit its structure. Indeed, that an interpretation I_i for the provision n is obligatory means that I_i is admissible and that no other (non-conflicting) interpretations for n is admissible.

Definition 11 (Obligatory A-Interpretations). *The proof condition of* defeasible provability for obligatory A-interpretations *is*

$+\partial^I_{\mathsf{Obl}}$*: If* $P(n+1) = +\partial^I_{\mathsf{Obl}} \pm \mathsf{I}_i(n,a)$ *then*
(1) $\mathsf{Obl} \pm \mathsf{I}_i(n,a) \in F$ *or*
  (2.1) $\sim\mathsf{Obl} \pm \mathsf{I}_i(n,a) \notin F$*, and*
  (2.2) $+\partial^I_{\mathsf{Adm}} \pm \mathsf{I}_i(n,a) \in P(1..n)$*, and*
  (2.3) $\forall s \in R[\pm\mathsf{I}_m(n,b), j]$*, such that* $\mathsf{I}_m(n,b) \neq \sim\mathsf{I}_i(n,a)$*, either*
    (2.3.1) s *is discarded for* $\pm\mathsf{I}_m(n,b)$*, or*
    (2.3.2) $\exists t \in R_{defT}[\sim \pm \mathsf{I}_m(n,b), k]$ *such that* t *is applicable for* $\sim \pm \mathsf{I}_m(n,b)$ *and* $t > s$.

The negative proof condition for obligatory interpretations is as follows.

Definition 12 (Non-obligatory A-Int.). *The proof condition of* defeasible refutability for obligatory A-interpretations *is*

$-\partial^I_{\mathsf{Obl}}$*: If* $P(n+1) = -\partial^I_{\mathsf{Obl}} \pm \mathsf{I}_i(n,a)$ *then*
(1) $\mathsf{Obl} \pm \mathsf{I}_i(n,a) \notin F$ *and*
  (2.1) $\sim\mathsf{Obl} \pm \mathsf{I}_i(n,a) \in F$*, or*
  (2.2) $-\partial^I_{\mathsf{Adm}} \pm \mathsf{I}_i(n,a) \in P(1..n)$*, or*
  (2.3) $\exists s \in R[\pm\mathsf{I}_m(n,b), j]$ *such that* $\mathsf{I}_m(n,b) \neq \sim\mathsf{I}_i(n,a)$ *and*
    (2.3.1) s *is applicable for* $\pm\mathsf{I}_m(n,b)$*, and*
    (2.3.2) $\forall t \in R[\sim \pm \mathsf{I}_m(n,b), k]$ *either* t *is discarded for* $\sim \pm \mathsf{I}_m(n,b)$ *or* $t \not> s$.

Example 4. *Consider the theory in Example 3. Facts make rules* r_5 *and* r_6 *applicable. Rule* r_7 *has an empty antecedent, so it is applicable, too. Despite* $r_8 \in R$*,* $b \not\sqsubseteq a$*, then* r_5 *and* r_6 *are in conflict. The theory assumes that* r_6 *is stronger than* r_5*, thus we would obtain* $+\partial^I_{\mathsf{Adm}}\mathsf{I}_1(art.575, a)$ *(and so* $-\partial^I_{\mathsf{Adm}}\mathsf{I}_2(art.575, b)$ *and* $-\partial^I_{\mathsf{Adm}}\mathsf{I}_3(art.575, b)$*). However, these last conclusions are not obtained because of* r_7*:* r_7 *and* r_6 *attack each other, thus we in fact have* $-\partial^I_{\mathsf{Adm}}\mathsf{I}_1(art.575, a)$*,* $-\partial^I_{\mathsf{Adm}}\mathsf{I}_1(art.575, a')$*, and also* $-\partial^I_{\mathsf{Adm}}\mathsf{I}_4(art.575, a)$*. Hence, we reinstate* $+\partial^I_{\mathsf{Adm}}\mathsf{I}_2(art.575, b)$ *via* r_5*. What interpretations are obligatory? Trivially, we get* $+\partial^I_{\mathsf{Obl}}\mathsf{I}_3(art.3,c)$*. Also, since* $\mathsf{I}_2(art.575, b)$ *is the only admissible interpretation of art. 575, then* $+\partial^I_{\mathsf{Obl}}\mathsf{I}_2(art.575, b)$ *(check conditions 2.2) and 2.3) in Definition 11). All other interpretations are refuted as obligatory.*

2.4 Proof Theory: Interpretation Outcomes

Let us show how to prove O-interpretations (i.e., modal literals following from A-interpretations): for example, if $\mathsf{I}_i(n,a)$ is admissible, then a should be admissible as well.

Definition 13 (Admissible O-Interpretations). *The proof condition of* defeasible provability for admissible O-interpretations *is*

$+\partial^{\mathsf{Adm}}$*: If* $P(n+1) = +\partial^{\mathsf{Adm}} l$ *then*
(1) $\mathsf{Adm}l \in F$ *or* $\mathsf{Obl}l \in F$*, or*
(2) $\exists \mathsf{I}_i \in \mathrm{Int}, \exists n \in \mathrm{NORM} : +\partial^I_{\mathsf{Adm}}\mathsf{I}_i(n,l) \in P(1..n)$.

Definition 14 (Refuted O-Interpretations). *The proof condition of* defeasible refutability for admissible O-interpretations *is*

$-\partial^{\mathsf{Adm}}$*: If* $P(n+1) = -\partial^{\mathsf{Adm}} l$ *then*
(1) $\mathsf{Adm}l \notin F$ *and* $\mathsf{Obl}l \notin F$*, and*
(2) $\forall \mathsf{I}_i \in \mathrm{Int}, \forall n \in \mathrm{NORM} : -\partial^I_{\mathsf{Adm}}\mathsf{I}_i(n,l) \in P(1..n)$.

Definition 15 (Obligatory O-Interpretations). *The proof condition of* defeasible provability for obligatory O-interpretations *is*

$+\partial^{\mathsf{Obl}}$*: If* $P(n+1) = +\partial^{\mathsf{Obl}} l$ *then*
(1) $\mathsf{Obl}l \in F$*, or*
(2) $\exists n \in \mathrm{NORM}$ *such that*
  (2.1) $\exists \mathsf{I}_i \in \mathrm{Int} : +\partial^I_{\mathsf{Adm}}\mathsf{I}_i(n,a) \in P(1..n)$ *and*
  (2.2) $\forall \mathsf{I}_j \in \mathrm{Int}, -\partial^I_{\mathsf{Adm}}\mathsf{I}_i(n,x) \in P(1..n)$ *if* $x \neq a$.

Definition 16 (Non-obligatory O-Int.). *The proof condition of* defeasible refutability for obligatory O-interpretations *is*

$-\partial^{\mathsf{Obl}}$*: If* $P(n+1) = -\partial^{\mathsf{Obl}} l$ *then*
(1) $\mathsf{Obl}l \notin F$*, and*
(2) $\forall n \in \mathrm{NORM}$:
  (2.1) $\forall \mathsf{I}_i \in \mathrm{Int} : -\partial^I_{\mathsf{Adm}}\mathsf{I}_i(n,a) \in P(1..n)$ *or*
  (2.2) $\exists \mathsf{I}_j \in \mathrm{Int}, +\partial^I_{\mathsf{Adm}}\mathsf{I}_i(n,x) \in P(1..n)$ *and* $x \neq a$.

Proof theory for O-interpretations is simple. Just notice that obligatory interpretations are nothing but literals resulting from obligatory A-interpretations.

EXAMPLE 5. *If we work again on Example 3, the proofs informally described in Example 4 show that* $+\partial^{\Box} b$, *where* $\Box \in \{\mathsf{Adm}, \mathsf{Obl}\}$.

3. OPTION 2: STRUCTURED NORMS

The second option is a refinement of the one presented in Section 2. We add deontic rules expressing legal provisions and interpretation rules here work on the literals composing those deontic rules.

More precisely, as in [8], we introduce obligation rules of the form $r : a_1, \ldots, a_n \Rightarrow^{\mathsf{Obl}} b$ to represent deontic legal provisions; if r is applicable and undefeated, then we can derive $\mathsf{Obl}b$. These rules represent in our framework the legal provisions on which interpretation options apply [5].

Second, interpretation rules work on the literals occurring in obligation rules. In other words, given a rule $r : a_1, \ldots, a_n \Rightarrow^{\mathsf{Obl}} b$, an interpretation function maps for each provision n the sequence $x = \langle a_1, \ldots, a_n, b\rangle$ of literals in r onto another sequence y of literals that can be identical (literal interpretation), partially different or completely different from x. Hence, an interpretation I_i is meant to make the original version of rule r unusable and the new one—where the literals are changed according to y—usable to derive an obligation. For instance, if $r : a_1, a_2 \Rightarrow^{\mathsf{Obl}} b$ and the interpretation I_i returns $y = \langle a_1, a_2', b'\rangle$, the *interpreted version* of r according to I_i is $r : a_1, a_2' \Rightarrow^{\mathsf{Obl}} b'$.

Let us extend and modify language and proof theory.

DEFINITION 17 (LANGUAGE, OPTION 2). *Let* $\mathrm{PROP} = \{a, b, \ldots\}$ *be a set of propositional atoms,* $\mathrm{NORM} = \{n_1, n_2, \ldots\}$ *a set of provision labels,* $\mathsf{INTR} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$ *a set of interpretation functions,* $\mathrm{MOD} = \{\mathsf{Obl}, \mathsf{Adm}\}$ *a set of modal operators, and* $\sqsubseteq$ *and* $\approx$ *the relations denoting semantic inclusion and equivalence.*

- *The set* Lit*, the complementary of literals, the set* ModLit*, the set* ModIntr*, the complementary of modal literals and of qualified interpretations are as in Definition 1;*
- *The set* $\mathrm{Int} = \{\mathsf{I}_i(n, x, y), \neg\mathsf{I}_i(n, x, y) | \exists \mathcal{I}_i : \mathrm{NORM} \times X \mapsto Y \in \mathsf{INTR} : X, Y \in \bigcup_{j=1}^{n} \mathcal{P}(\mathrm{Lit}^n)$ *and* $x \in X, y \in Y\}$ *denotes the set of* interpretations *and their negations: an expression* $\mathsf{I}_i(n, x, y)$*, for instance, means that the interpretation* I_i *of provision* n *(corresponding to the sequence of literals* x*) returns a tuple* y *of literals that should instead occur in* n*.*
- *The* complementary of an interpretation ϕ *is denoted by* $\sim\phi$ *and is defined as follows:*

ϕ	$\sim\phi$
$\mathsf{I}_i(n, x, y)$	$\sim\mathsf{I}_i(n, x, y) \in \{\neg\mathsf{I}_i(n, x, y), \mathsf{I}_i(n, x, z), \mathsf{I}_j(n, x, w) \mid \exists a$ *in* y, $\exists b$ *in* w : *either* $a \not\sqsubseteq b$ *or* $b \not\sqsubseteq a\}$
$\neg\mathsf{I}_i(n, x, y)$	$\sim\neg\mathsf{I}_i(n, x, y) \in \{\mathsf{I}_i(n, x, y), \mathsf{I}_i(n, x, w) \mid \exists a$ *in* y, $\exists b$ *in* $w : a \approx b\}$.

Let us define the set of rules, which consists of the set of meaning rules, obligation rules, and interpretation rules.

Meaning rules (i.e., strict rules with the arrow $\rightarrow^{M}$) are exactly as in Section 2 and their definition is omitted.

DEFINITION 18 (OBLIGATION RULES). *Let* Lab *be a set of arbitrary provision labels. The set* $\mathrm{Rul}^{\mathsf{Obl}}$ *of* obligation rules *consists of rules of the following type*

$$r : A(r) \hookrightarrow^{\mathsf{Obl}} C(r)$$

where

1. $r \in \mathrm{Lab}$ *is the name of the rule, and*
2. $A(r) = \{a_1, \ldots, a_n\}$*, the* antecedent *(or* body*) of the rule is such that each* $a_i \in A(r)$ *is a literal* $l \in \mathrm{Lit}$*;*
3. $\hookrightarrow \in \{\Rightarrow, \rightsquigarrow\}$*;*
4. $C(r) = l$ *is the* consequent *(or* head*) of the rule such that* $l \in \mathrm{Lit}$*.*

DEFINITION 19 (⊗-EXPRESSIONS, OPTION 2). *⊗-expressions are defined as follows:*

1. *Each interpretation* $\mathsf{I}_i(n, x, y)$ *or* $\neg\mathsf{I}_i(n, x, y)$ *is an ⊗-expressions where*
   (a) n *is an obligation rule* $a_1, \ldots, a_n \hookrightarrow^{\mathsf{Obl}} b$*;*
   (b) $x = \langle a_1, \ldots, a_n, b\rangle$ *and* $y = \langle a_1', \ldots, a_n', b'\rangle$*;*
2. *If* $\phi_1, \ldots, \phi_n$ *are interpretations, then* $\phi_1 \otimes \cdots \otimes \phi_n$ *is an ⊗-expression;*
3. *Nothing else is an ⊗-expression.*

REMARK 2. *We state that interpretations work on* n*-tuples of literals returning other* n*-tuples. This option makes things simple: each literal can be possibly replaced but not removed, and no new literals can be added. Hence, we cannot directly express the shift from man* $\Rightarrow^{\mathsf{Obl}}$ *b into adult, male* $\Rightarrow^{\mathsf{Obl}}$ *b. To cope with this, it suffices to state that any* $\mathsf{I}_i(n, x, y)$ *is such* x *and* y *have the form* $\langle X, l\rangle$ *where* X *is a set of literals. Conflicting interpretations leading to* $\langle X, l\rangle$ *and* $\langle Y, l'\rangle$ *would require to check if* $X \neq Y$ *or* $l \neq l'$*.*

Interpretation rules (i.e., defeasible rules with the arrow $\Rightarrow^{I}$) are exactly as in Section 2: the difference is here that their consequents are ⊗-expressions as in Definition 19. Interpretation theories are as in Definition 3.

3.1 Proof Theory

Proof theory must be significantly revised. A first major change is that we can derive obligatory literals using obligation rules (i.e., legal provisions). A second major change is that the derivation of an obligatory interpretation with respect to a provision n has a double effect: (i) making ineffective the original version of n; (ii) making applicable the new version that follows from the derived obligatory interpretation. In other words, the provability of interpretations allows us to update the set of applicable obligation rules: obligatory interpretations thus change obligation rules and work in a similar way of legal modifications as modelled in [9]. This means that we need a mechanism for deriving *usable obligation rules*, i.e., those obligation rules that can be used to derive obligations: if the interpretation $\mathsf{I}_i(r, \langle a_1, a_2, b\rangle, \langle a_1, a_2', b'\rangle)$ is derived (as obligatory), then $r : a_1, a_2 \Rightarrow^{\mathsf{Obl}} b$ is unusable and $r : a_1, a_2' \Rightarrow^{\mathsf{Obl}} b'$ is usable, i.e., this last version of r can only be employed to derive obligations. In contrast, deriving A-interpretations does not substantially change: the only difference is in the format of

interpretation rules ($\otimes$-expressions are different, and modal literals can be differently proved), but proofs run exactly as before. It is enough to keep Definitions 9, 10, 11, and 12.

Definition 20. *An obligation rule $r \in R^{\mathsf{Obl}}$ is* applicable *if it is body-applicable. An obligation rule $r \in R^{\mathsf{Obl}}$ is* discarded *if it is body-discarded.*

Let us first state the procedures for deriving usable obligation rules, which require to integrate Definition 4 with new proof tags $\pm\partial^R$.

Definition 21 (Usable obligation rules). *The proof condition of* defeasible provability for usable obligation rules *is*
$+\partial^R$*: If* $P(n+1) = +\partial^R r : a_1, \ldots, a_n \hookrightarrow^{\mathsf{Obl}} b$ *then*
(1) $r \in R^{\mathsf{Obl}}$*, and, either*
(2) if $A(r) = \{a_1, \ldots, a_n\}$ *and* $C(r) = b$*, then*
$\forall \mathsf{I}_i \in \mathrm{Int} : -\partial^I_{\mathsf{Adm}}\mathsf{I}_i(r, x, y)$ *if* $y \neq \langle a_1, \ldots, a_n, b\rangle$*, or*
(3) $+\partial^I_{\mathsf{Obl}}\mathsf{I}_i(r, x, y)$ *if* $y = \langle a_1, \ldots, a_n, b\rangle$*.*

Definition 22 (Unusable obligation rules).
$-\partial^R$*: If* $P(n+1) = -\partial^R r : a_1, \ldots, a_n \hookrightarrow^{\mathsf{Obl}} b$ *then*
(1) $r \notin R^{\mathsf{Obl}}$*, or*
(2) $A(r) = \{a_1, \ldots, a_n\}$*,* $C(r) = b$*, and*
$\exists \mathsf{I}_i \in \mathrm{Int} : +\partial^I_{\mathsf{Adm}}\mathsf{I}_i(r, x, y)$ *and* $y \neq \langle a_1, \ldots, a_n, b\rangle$*, or*
(3) $-\partial^I_{\mathsf{Obl}}\mathsf{I}_i(r, x, y)$ *and* $y = \langle a_1, \ldots, a_n, b\rangle$*.*

Remark 3. *An obligation rule $r : a_1, \ldots, a_n \Rightarrow^{\mathsf{Obl}} b$ is usable if one of the following mutually exclusive conditions holds: either (2) $r : a_1, \ldots, a_n \Rightarrow^{\mathsf{Obl}} b$ is in the set R of rules of the theory and no other version is admissible, or (3) a different version is in R but an interpretation I_i is provable as obligatory that transforms this version into the one where $\{a_1, \ldots, a_n\}$ and $C(r) = b$. Notice that we could relax condition (3) and just require that I_i is admissible: we would change proof theory in such a way as to possibly get more derived compatible versions of r, i.e., that the same provision n may lead to different normative sentences that are implied by n.*

Definition 23 (Provability for obligations).
The proof condition of defeasible provability for obligations *is*
$+\partial^{\mathsf{Obl}}$*: If* $P(n+1) = +\partial^{\mathsf{Obl}} l$ *then*
(1) $\mathsf{Obl} l \in F$*, or*
(2.1) $\sim\mathsf{Obl} l \notin F$ *and*
(2.2) $\exists r \in \mathrm{Rul}^{\mathsf{Obl}}[l] : +\partial^R r \in P(1..n)$*, r is applicable, and*
(2.3) $\forall s \in R^{\mathsf{Obl}}[\sim l]$*, either*
(2.3.1)s is discarded, or
(2.3.2) $\exists t \in R^{\mathsf{Obl}}[l] : t$ *is applicable and* $t > s$*.*

Definition 24 (Refutability for obligations).
The proof condition of defeasible refutability for obligations *is*
$-\partial^{\mathsf{Obl}}$*: If* $P(n+1) = -\partial^{\mathsf{Obl}} l$ *then*
(1) $\mathsf{Obl} l \notin F$*, and*
(2.1) $\sim\mathsf{Obl} l \in F$ *or*
(2.2) $\forall r \in \mathrm{Rul}^{\mathsf{Obl}}[l] : -\partial^R r \in P(1..n)$*, or r is discarded, or*
(2.3) $\exists s \in R^{\mathsf{Obl}}[\sim l]$:
(2.3.1)s is applicable, and
(2.3.2) $\forall t \in R^{\mathsf{Obl}}[l]$: *either t is discarded or* $t \not> s$*.*

Remark 4. *The conditions for proving obligations are very similar to the ones in [8]. A relevant difference is that we require here that any rule employed to derive an obligation must be usable. Interestingly, the provability of obligations affects the derivation of interpretations. Consider when, given an obligation rule $r : a \Rightarrow^{\mathsf{Obl}} b$ in R, an interpretation rule $r' : \mathsf{Obl} b \Rightarrow^I \mathsf{I}_i(r, \langle a, \neg b\rangle)$ holds, which would lead to $r : a \Rightarrow^{\mathsf{Obl}} \neg b$, if I_i results from r' to be obligatory. As expected, we cannot prove this new version of r, whose provability relies on an incompatible reading of r.*

Example 6. *Let us develop Example 3 introducing the new ideas of this section. Consider the following:*[9]

$$\begin{array}{lcl} a & = & \langle kill\_man, 21y \leq\rangle \\ a' & = & \langle kill\_male, 21y \leq\rangle \\ b & = & \langle kill\_person, 21y \leq\rangle \\ c & = & \langle person, equal\_status\rangle \\ c' & = & \langle citizen, equal\_status\rangle \\ d & = & \langle mother, kill\_newborn, \neg 21y \leq\rangle \end{array}$$

$F = \{kill\_adult, kill\_female, \mathsf{Obl}\,\mathsf{I}_3(art.3, c', c),$
$\mathsf{Obl}\,\mathsf{I}_1(art.578, d, d), \neg capable\}$
$R = \{r_5 : kill\_adult, kill\_female, \mathsf{Obl}\,\mathsf{I}_3(art.3, c', c) \Rightarrow^I$
$\mathsf{I}_2(art.575, a, b) \otimes \mathsf{I}_3(art.575, a, b) \otimes \neg\mathsf{I}_1(art.575, a, a),$
$r_6 : \mathsf{Obl}\,\mathsf{I}_1(art.578, d, d) \Rightarrow^I \mathsf{I}_1(art.575, a, a) \otimes \mathsf{I}_4(art.575, a, a),$
$r_7 :\Rightarrow^I \mathsf{I}_1(art.575, a, a'),$
$r_8 : b \rightarrow^M a,\ r_9 : kill\_female \rightarrow^M kill\_person,$
$art.575 : kill\_man \Rightarrow^{\mathsf{Obl}} 21y \leq,$
$art.85 : \neg capable \rightsquigarrow^{\mathsf{Obl}} \neg 21y \leq\}$
$>= \{r_6 > r_5\}$

The theory above extends the one in Example 3 by adding (i) r_9, which clarifies that killing any female human is killing a person, (ii) art.575, which is the obligation rule encoding the logical structure of art. 575 as stated in the Italian penal code, (iii) art.85, which exhibits the logical content of art. 85 of the Italian penal code, a provision that prevents to prove that one is punishable for homicide if incapable.

Proofs of A- and O-interpretations are exactly as explained in Example 4. The novelty is that provability of interpretations may lead to new usable versions of a same provision, thus affecting what obligatory literals may hold. Trivially, since $\mathsf{Obl}\,\mathsf{I}_3(art.3, c', c)$ and $\mathsf{Obl}\,\mathsf{I}_1(art.578, d, d)$ are in R, we have that $+\partial^R art.3 : person \Rightarrow^{\mathsf{Obl}} equal\_status$ and $+\partial^R art.578 : mother, kill\_newborn \Rightarrow^{\mathsf{Obl}} \neg 21y \leq$. Also, (a) $+\partial^I_{\mathsf{Adm}}\mathsf{I}_2(art.575, a, b)$, which makes unusable the original version of $art.575 \in R$, and (b) $+\partial^I_{\mathsf{Obl}}\mathsf{I}_2(art.575, a, b)$, which supports the provability of $+\partial^R art.575 : kill\_person \Rightarrow^{\mathsf{Obl}} 21y \leq$. The new version of art. 575 is usable: since we obtain $+\Delta^M kill\_person$ from $kill\_female$ via r_9, we would have an argument supporting the derivation of $+\partial^{\mathsf{Obl}} 21y \leq$ from $kill\_female$ (which is not possible with the original version of art. 575). However, this is not the case, because art. 85 is applicable and attacks the conclusion of art. 575, thus leading to $-\partial^{\mathsf{Obl}} 21y \leq$.

[9]For the sake of simplicity, let us assume that the consequent of art. 578 is $\neg 21y \leq$ (at least 21 years prison), which is implied by any imprisonment between 4 and 12 years.

4. SUMMARY AND DISCUSSION

This paper presented a logical machinery for reasoning about interpretive canons, which is based on the following intuitions: (a) canons are represented by defeasible rules; (b) different reasoning patterns can be identified depending on whether we work on interpretations as activities or as outcomes [13]; (c) competing interpretive options can be handled by stating a priority over conflicting rules, but different ranking preferences can also be introduced among compatible interpretive acts; (d) canons are defeasible rules licensing deontic interpretive claims; (e) the logic can deal with the interpretation of abstract, non-analysed provisions and of structured provisions.

It is not hard to show that the proposed framework enjoys some relevant properties (see [7]).

DEFINITION 25. *An Interpretation Theory $D = (F, R, >)$ is* consistent *iff $>$ is acyclic and F does not contain pairs of complementary (modal) literals and interpretations.*

PROPOSITION 26. *Let D be a consistent Interpretation Theory, and $\Box \in$ MOD. For any interpretation ϕ or literal l, it is not possible to have (a) both $D \vdash +\partial^I_\Box \phi$ and $D \vdash -\partial^I_\Box \phi$, (b) $D \vdash +\partial^\Box l$ and $D \vdash -\partial^\Box l$, (c) both $D \vdash +\partial^I_\Box \phi$ and $D \vdash +\partial^I_\Box \sim\phi$, (d) $D \vdash +\partial^\Box l$ and $D \vdash +\partial^\Box \sim l$.*

PROPOSITION 27. *Let D be a consistent Interpretation Theory. For any A- and O-interpretation ϕ and $\Box \in$ MOD:*

1. *if $D \vdash +\partial^I_{\mathsf{Obl}} \phi$, then $D \vdash -\partial^I_\Box \sim\phi$;*
2. *if $D \vdash +\partial^I_{\mathsf{Adm}} \phi$, then $D \vdash -\partial^I_{\mathsf{Obl}} \sim\phi$;*
3. *$D \vdash +\partial^I_{\mathsf{Obl}} \phi$, then $D \vdash +\partial^I_{\mathsf{Adm}} \phi$;*
4. *if $D \vdash +\partial^I_{\mathsf{Obl}} \mathsf{I}_i(n, a)$, then $D \vdash -\partial^I_{\mathsf{Adm}} \mathsf{I}_j(n, a')$;*[10]
5. *if $D \vdash +\partial^I_\Box \mathsf{I}_i(n, a)$, then $D \vdash +\partial^\Box a$.*

If we assume that, given an interpretive theory D, the set of conflicting interpretation rules has been defined, the following holds for Option 1:[11]

THEOREM 28 (COMPLEXITY, OPTION 1; CF. [7]). *Let $D = (F, R >)$ be an Interpretation Theory and U^D be the set of all the atomic literals and atomic A-interpretations occurring in D. The set of conclusions of D can be computed in time linear to the size of the theory, i.e., $O(|R| * |U^D|)$.*

We leave to future research to study the complexity of Option 2, where usable obligation rules can be proved.

Finally, a valuable aspect of the proposed machinery is that it can accommodate different doctrinal views regarding legal interpretation. In particular, we argued that two different but non-conflicting interpretations of the same provision can be admissible. First of all, we must notice that a number of logical relations can hold between meanings [3, 4]. All these aspects can be modularly introduced in the language definition. Also, one may argue that any interpretations $\mathsf{I}_i(n, a)$ and $\mathsf{I}_j(n, a')$ of n—if we just consider Option 1 of this paper—are in conflict [14] whenever $a \neq a'$, irrespective of any possible logical relation (entailment, semantic overlapping, etc.) between a and a'. Suppose we accept this last view. Then, we would simply have to amend condition 5 of Definition 1 and modularly modify proof theory: if $\mathsf{I}_i(n, a)$ and $\mathsf{I}_j(n, a')$ are always in conflict, then Definitions 11-12 and 15-16 no longer apply. If so, a different way of proving as obligatory an A- and O-interpretation ϕ could be based on checking whether ϕ is provable for any set F of facts in the theory or it rather depends on specific facts: in the first case ϕ would be obligatory, in the second we would just have a case of admissibility. A systematic investigation on these options is left for the future research.

[10] Similar results hold for Option 2.

[11] The relation $\sqsubseteq$ requires to compute the monotonic part of the theory using, as facts, all sets in the powerset of the set of literals occurring in the antecedents of meaning rules.

References

[1] R. Alexy and R. Dreier. Statutory interpretation in the Federal Republic of Germany. In MacCormick and Summers [11].

[2] G. Antoniou, D. Billington, G. Governatori, and M. Maher. Representation results for defeasible logic. *ACM Trans. Comput. Log.*, 2(2):255–287, 2001.

[3] M. Araszkiewicz. Towards systematic research on statutory interpretation in AI and law. In *Proc. JURIX 2013*. IOS Press, 2013.

[4] M. Araszkiewicz. Scientia juris: A missing link in the modelling of statutory reasoning. In *Proc. JURIX 2014*. IOS Press, 2014.

[5] G. Boella, G. Governatori, A. Rotolo, and L. van der Torre. A logical understanding of legal interpretation. In *Proc. KR 2010*. AAAI Press, 2010.

[6] B. Brozek. Legal interpretation and coherence. In M. Araszkiewicz and J. Savelka, editors, *Coherence: Insights from Philosophy, Jurisprudence and Artificial Intelligence*. Springer, 2013.

[7] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco. Computing strong and weak permissions in defeasible logic. *J. Philosophical Logic*, 42(6):799–829, 2013.

[8] G. Governatori and A. Rotolo. BIO logical agents: Norms, beliefs, intentions in defeasible logic. *Auton. Agent Multi Agent Syst.*, 17(1):36–69, 2008.

[9] G. Governatori and A. Rotolo. Changing legal systems: Legal abrogations and annulments in defeasible logic. *Logic Journal of the IGPL*, 18:157–194, 2010.

[10] F. Macagno, G. Sartor, and D. Walton. Argumentation schemes for statutory interpretation. In *Proc. ARGUMENTATION 2012*. Masaryk University, 2012.

[11] D. MacCormick and R. Summers, editors. *Interpreting Statutes: A Comparative Study*. Ashgate, 1991.

[12] H. Prakken and G. Sartor. Formalising arguments about norms. In *Proc. JURIX 2013*. IOS Press, 2013.

[13] A. Ross. *On Law and Justice*. Stevens, London, 1958.

[14] G. Sartor, D. Walton, F. Macagno, and A. Rotolo. Argumentation schemes for statutory interpretation: A logical analysis. In *Proc. JURIX 2014*. IOS Press, 2014.

[15] G. Tarello. *L'interpretazione della legge*. Giuffrè, 1980.