Deontic Explanations in Åqvist's Systems

Agata Ciabattoni, Blaž Istenič Urh, Xavier Parent [1]

Vienna University of Technology (TU Wien)

Abstract

We propose a logic-based formalization of deontic explanations, which aim to address questions about obligations and permissions such as "Why does this norm apply?" or "Why should I perform A rather than B?". Our framework is designed to be flexible, accommodating various deontic logics. In this work, we build on Åqvist's dyadic deontic logics, extending them by introducing an obligation operator that supports Factual Detachment along with mechanisms for non-monotonic reasoning. Applying our framework to legal examples, we demonstrate its effectiveness in explaining normative statements while also identifying its limitations.

Keywords: Norms, Deontic explanations, Åqvist's systems, factual detachment, specificity, Contrary-To-Duties

1 Introduction

Explanations answer "why?" rather than just "what?" questions, helping us understand the reasons behind a claim rather than taking it at face value. They are essential in everyday life, aiding decision-making, justifying our decisions, and communicating our reasoning to others. In Artificial Intelligence, explanations enhance trust by clarifying why a system reaches a particular decision.

Explanations clarify not just what is true but also why it is so. This applies not only to factual beliefs, but also to the norms –obligations, prohibitions, and permissions– that shape our lives. Understanding whether a norm applies is important, but so is knowing why. For example, the GDPR grants individuals the right to be forgotten, yet legal obligations may require to retain certain data (e.g., for tax or audit purposes). Why should then the data be erased or retained in a particular situation? These types of inquiries yield answers known as *deontic explanations*. The central topic of this paper is their formalization and analysis through logic. In our context, explanations serve to constrain the

[1] Work supported by the Austrian Science Fund (FWF) and the Luxembourg National Research Fund (FNR) through the LoDEX project [doi: 10.55776/I6372 and INTER/DFG/23/17415164/LODEX], the FWF project LFforGDR [doi: 10.55776/PAT2141924], and by the Vienna Science and Technology Fund (WWTF) Grant ID 10.47379/ICT23030.

application of a normative system, ensuring that the reasoning process remains transparent and contextually relevant.

We present a framework for deontic explanations that (i) accounts for the unique aspects of normative reasoning and (ii) is flexible enough to accommodate different deontic logics. To illustrate its application, we use legal examples.

Ad (i). Reasoning about norm violation and nonmonotonic reasoning are often regarded (to hijack a famous title) as the two "fundamental problem[s] of deontic logic" [24]. In the past, they have mostly been studied independently of one another. There is a pressing need for an explanatory framework that effectively integrates both aspects. Our central deontic concept is what we, following [42], call a "deliberative" obligation. This notion includes a mechanism for handling the overriding of obligations based on specificity. [2] An analogous concept for permission is also introduced and similar considerations are used to define a notion of contrastive explanation, where both obligations and permissions can explicitly interact.

Ad (ii). In this paper, we instantiate the framework using Åqvist's systems **E**, **F**, and **G** for conditional obligation. However, other logics for conditional obligation could, in principle, be used as well. Broadly speaking, existing deontic logics fall into two categories: preference-based and norm-based systems [14,13]. Åqvist's systems are landmark representatives of the former; recent developments, such as the introduction of sequent-style calculi [8,9,10] and encodings into classical logic [36], further motivate our choice to work with them. In Åqvist's systems, modalities are analyzed within possible world semantics, where a preference relation orders possible worlds according to their betterness. These systems were first introduced to address the paradox of Contrary-To-Duty (CTD) reasoning [7] (norms that come into force when other norms are violated). The semantics is accompanied by an intuitive graphical representation using "layers", making it easier to convey the concepts to non-experts. We present a novel solution within this framework to address the well-known problem posed by various forms of detachment, specifically the so-called (strong) factual vs deontic detachment. We contend that the way detachment is handled in Åqvist's systems is inadequate for explanatory purposes. To resolve this issue, we introduce a monadic obligation operator that has factual detachment built-in. The proposed solution is sufficiently general and can be easily applied to logics of conditional obligation with different semantics (e.g., those employing a selection function [6]).

To illustrate further the framework, we analyze a legal case involving generalized quantifiers such as "most". This analysis is conducted using the extension of Åqvist's systems from [29]. The resulting explanations, emerging from the interaction between deontic and normality conditionals, are well-supported and give further credibility to the proposed approach.

The paper is structured as follows: Section 2 recalls Åqvist's systems, which form the basis of our work. Section 3 introduces a general definition of deontic

[2] More specific norms override more generic ones.

explanation and illustrates its application through legal examples. Section 4 tackles the detachment problem and presents our solution using a GDPR-based case study. Section 5 explores deliberative obligation and permission in the context of deontic explanations. Section 6 discusses contrastive explanation, while Section 7 examines a legal case where normality plays a crucial role.

Related works: There is a vast literature on explanation in the empirical sciences, beginning with Hempel's seminal work [20], as well as on the nature of action explanations, following Davidson [11]. In Artificial Intelligence, research has flourished since [27]. To our knowledge, the notion of deontic explanation has been mainly analyzed from the argumentation perspective; recent works include [12,16,35,44]. The work [35] examines various explanatory cases, including legal interpretation, within argumentation frameworks with structured arguments. In our paper, we assume instead a set of norms and focus on their derivability without explicitly modeling norm interactions. The approach [44] integrates argumentation theory with input/output logics. These logics belong to the norm-based tradition of deontic logic, as opposed to preference-based approaches, where Åqvist systems are prominent. The focus there is on contrastive explanations arising from dialogues, making the explananda more complex. To compare, our contrastive explanations have a counterfactual flavor, identifying the difference-making facts that determine which norm holds over another. The notion of stable normative explanation has been developed in [16] and [12], incorporating non-monotonic and argumentative elements. Like our approach, their explanations are sets of formulas. However, their primary focus is on stability—ensuring that explanations remain valid even when new facts are introduced, which is not our main concern, as we focus on explanations derived from a fixed knowledge base (and not from its possible updates). Another key difference is that [16] and [12] allow for prioritization among norms to resolve conflicts, whereas we leave them be. Finally, while we incorporate some non-monotonic elements in our notions of deliberative obligation and contrastive explanation, Åqvist's systems – without the normality conditional from [29] – remain monotonic, a notable difference with the above approaches.

Note that in argumentation theory, normative concepts are not given a truth-functional semantics. This lack of semantics presents a challenge for explanatory purposes, as counter-models are often the most effective (if not the only) way to demonstrate that an obligation does not hold. Moreover, logic-based approaches may offer advantages in automation—for instance, Åqvist's systems can be encoded into SAT solvers, as shown in [36].

2 Preliminaries on Åqvist's systems

The language $\mathcal{L}$ of Åqvist's systems **E**, **F**, and **G** [47] is defined by the grammar:

$$A ::= p \in PropVar \mid \neg A \mid A \wedge A \mid \Box A \mid \bigcirc(A/A)$$

$\bigcirc(A/B)$ is read as *given B, A is obligatory*, and $\Box A$ as *A is settled (as true)*. The former pertains to norms (is deontic) and the latter to facts and truth (is alethic). The other Boolean connectives, $\top$, $\bot$, and $\Diamond$ are defined as usual, the

permission operator as $P(A/B) ::= \neg\bigcirc(\neg A/B)$. $\bigcirc A$ and PA are short for $\bigcirc(A/\top)$ and $P(A/\top)$, respectively.

The semantics is defined in terms of preference models.

Definition 2.1 [Preference model] A preference model $\mathfrak{M} = \langle W, \succeq, V\rangle$ is a tuple consisting of a non-empty set of worlds W, a (preference) relation on the set of worlds $\succeq \subseteq W \times W$, and of a valuation function $V : \mathit{Prop Var} \mapsto \mathcal{P}(W)$, assigning to each propositional letter p the set of worlds where p holds. Intuitively, $w \succeq v$ is read as "w is at least as good as v". $w \succ v$ is read as "w is strictly better than v" and is defined by $w \succeq v$ and $v \not\succeq w$.

Boolean connectives are interpreted as usual; for the modalities, we have:

$$\mathfrak{M}, w \models \Box A \text{ iff } \forall v \in W\ \mathfrak{M}, v \models A$$
$$\mathfrak{M}, w \models \bigcirc(B/A) \text{ iff } \forall v \in \mathit{best}(A)\ \ \mathfrak{M}, v \models B$$

where $\mathit{best}(A) = \{v \in W \mid \mathfrak{M}, v \models A \text{ and for all } z \text{ s.t. } z \succeq v \text{ and } \mathfrak{M}, z \models A, v \succeq z\}$. Intuitively, $\bigcirc(B/A)$ holds if all the best A-worlds are B-worlds.

A formula A is *valid in a model* $\mathfrak{M}$ iff it is true at every world of $\mathfrak{M}$ and it is *valid* (tout court) iff it is valid in all models.

The systems differ based on the conditions imposed on $\succeq$. **E** requires $\succeq$ be reflexive: $\forall w \in W, w \succeq w$. **F** also demands limitedness:

$$\text{If } \exists w \in W \text{ s.t. } \mathfrak{M}, w \models A, \text{then } \mathit{best}(A) \neq \emptyset \qquad \text{(limitedness)}$$

G adds the requirements of transitivity and totality:

$$\text{If } \forall w, v, z \in W, \text{ if } w \succeq v \text{ and } v \succeq z, \text{ then } w \succeq z \qquad \text{(transitivity)}$$
$$\forall w, v \in W, \text{ either } w \succeq v \text{ or } v \succeq w \qquad \text{(totality)}$$

Definition 2.2 The axiomatization of **E** consists of any Hilbert system for classical propositional logic, the Modus Ponens rule (MP): If $\vdash A$ and $\vdash A \to B$, then $\vdash B$, the rule of Necessitation (Nec): If $\vdash A$, then $\vdash \Box A$, S5 axioms for $\Box$, and the following deontic axioms:

$$\bigcirc(B \to C/A) \to (\bigcirc(B/A) \to \bigcirc(C/A)) \qquad \text{(COK)}$$
$$\bigcirc(B/A) \to \Box\bigcirc(B/A) \qquad \text{(Abs)}$$
$$\Box A \to \bigcirc(A/B) \qquad \text{(Box-to-O)}$$
$$\Box(A \leftrightarrow B) \to (\bigcirc(C/A) \leftrightarrow \bigcirc(C/B)) \qquad \text{(Ext)}$$
$$\bigcirc(A/A) \qquad \text{(Id)}$$
$$\bigcirc(C/A \wedge B) \to \bigcirc(B \to C/A) \qquad \text{(Sh)}$$

F extends **E** with (D*), while **G** further adds (RM):

$$\Diamond A \to (\bigcirc(B/A) \to P(B/A)) \qquad \text{(D*)}$$
$$(P(B/A) \wedge \bigcirc(C/A)) \to \bigcirc(C/A \wedge B) \qquad \text{(RM)}$$

Sequent-style calculi for **E**, **F** and **G** were developed in [8,9,10], respectively, and an encoding into SAT solvers in [36].

Remark 2.3 The presence of axiom (D*) in **F** and **G** ensures that $\bigcirc(B/A) \wedge \Diamond A \vdash P(B/A)$, a property that does not hold in **E**.

Detachment

The principle of detachment (or modus ponens) converts a conditional obligation into an unconditional one, akin to discharging an assumption in natural deduction. Scholars concur that detachment is essential for reasoning. A system lacking this principle would have little practical value: without it, an agent will never un-conditionalize her obligations, and act according to what she believes to be best. Yet, scholars disagree on the precise formulation of this principle. Three distinct forms of detachment have been identified, each varying based on the modal status required of the antecedent:

$$\frac{\bigcirc(B/A) \qquad A}{\bigcirc B}\ \text{FD} \qquad \frac{\bigcirc(B/A) \qquad \Box A}{\bigcirc B}\ \text{SFD} \qquad \frac{\bigcirc(B/A) \qquad \bigcirc A}{\bigcirc B}\ \text{DD}$$

The abbreviations FD, SFD, and DD stand for Factual Detachment, Strong Factual Detachment, and Deontic Detachment, respectively. FD requires A to be true, SFD requires $\Box A$ to be true, and DD requires $\bigcirc A$ to be true.

These forms of detachment have primarily been discussed in the context of Contrary-to-Duty (CTD) scenarios, particularly Chisholm's paradox [7], which is often cited as evidence that FD and DD cannot coexist, see, e.g., [47, §9]. Notably, Åqvist's systems support only SFD and DD.

3 Deontic explanations

This section provides a general definition of deontic explanation and applies it to legal examples using Åqvist's systems.

First, we settle some terminology. We call *explanans* what is to be explained and *explanandum* what explains it. A *fact* is any formula that does not contain deontic modalities. A *norm* is a Boolean combination of formulas having a deontic modality, possibly negated, as the main connective. Finally, a *mixed formula* is any Boolean combination of norms (at least one), and facts.

Given a knowledge base $\mathcal{K}$ consisting of norms and facts in a deontic logic, a deontic explanation identifies the subset Γ (by abuse of notation, we will treat it both as a set of sentences and also their conjunction) that permits to obtain a mixed formula A. As we do not want the explanation to contain redundant elements, we start with the following definition (we use $\rhd$ for the consequence relation, without distinguishing between semantic and syntactic entailment):

Definition 3.1 [Minimal deontic explanation] Given a knowledge base $\mathcal{K}$, $\Gamma \subseteq \mathcal{K}$ minimally explains a mixed formula A iff

(i) $\Gamma \rhd A$ and

(ii) $\nexists \Gamma' \in \mathcal{K}$, s.t. $\Gamma \neq \Gamma'$, $\Gamma \rhd \Gamma'$, and $\Gamma' \rhd A$

The definition identifies a minimal set of reasons that entail A. We take both the explanandum (Γ) and the explanans (A) to be mixed formulas. For the explanandum, this is obvious: norms alone or combined with factual elements should be applied to obtain A. For A, the clearest case is when it is a norm,

but we also account for violations which are mixed formulas in Åqvist systems. A violation of the norm[3] $\bigcirc(C/B)$ is expressed as $\bigcirc(C/B) \wedge B \wedge \neg C$.

To illustrate the definition, take the following example. As in all our examples (except for Ex. 7.1), we use the language $\mathcal{L}$ of Åqvist systems and write $\rhd$ for their syntactic and semantic consequence relation.

Example 3.2 (Motor vehicle law 1) *Consider the following sentences, adapted and simplified from [40], §99, and [41], §1:*

(1) *In poor visibility as a result of adverse weather, it is compulsory to keep your lights on while driving,* $\bigcirc(l/w)$

(2) *In darkness, it is compulsory to keep your lights on while driving,* $\bigcirc(l/d)$

(3) *It is obligatory to pay a toll for using federal highways in Austria,* $\bigcirc(t/u)$

$\mathcal{K}^{V1} = \{\bigcirc(l/w), \bigcirc(l/d), \bigcirc(t/u)\}$.

(i) $\Gamma = \bigcirc(l/w) \wedge \bigcirc(l/d)$ *is the minimal explanation of* $\bigcirc(l/w \vee d)$. *Then, by monotonicity of Åqvist's systems, also* $\bigcirc(l/w) \wedge \bigcirc(l/d) \wedge \bigcirc(t/u) \rhd \bigcirc(l/w \vee d)$ *and so* $\Gamma = \bigcirc(l/w) \wedge \bigcirc(l/d) \wedge \bigcirc(t/u)$ *would explain* $\bigcirc(l/w \vee d)$. *However,* $\bigcirc(t/u)$ *is an irrelevant part of the explanation which is taken out by the minimality condition.*

(ii) Let us now see how we explain a violation. Take $\mathcal{K}^{V1} \cup \{w, \neg l\}$. *Then,* $\bigcirc(l/w) \wedge \bigcirc(l/d) \wedge w \wedge \neg l$ *minimally explains why the norm* $\bigcirc(l/w \vee d)$ *is violated (i.e.,* $\bigcirc(l/w \vee d) \wedge (w \vee d) \wedge \neg l$ *is entailed).*

A minimal deontic explanation provides a specific reason for why a given norm holds. However, the same knowledge base may contain multiple distinct minimal deontic explanations for the same mixed formula. In such cases, A is fully explained only by the disjunction of all minimal explanations, i.e.:

Definition 3.3 [(Complete) deontic explanation] Given a knowledge base $\mathcal{K}$ based on a deontic system DS, the disjunction of all $\Gamma \subseteq \mathcal{K}$, s.t. Γ minimally explains A, is the complete deontic explanation of A.

The example below illustrates the definition.

Example 3.4 (Motor vehicle law 2) *Take the following sentence, again adapted and simplified from [40], §99 Abs. 5:*

> *In poor visibility, it is obligatory to use low-beam headlights, front fog lights, or both,* $\bigcirc(l \vee f/\neg v)$, $P(l \wedge f/\neg v)$, $P(l \wedge \neg f/\neg v)$,$P(\neg l \wedge f/\neg v)$ [4]

This leads to: $\mathcal{K}_1^{V2} = \{\bigcirc(l \vee f/\neg v), P(l \wedge f/\neg v), P(l \wedge \neg f/\neg v), P(\neg l \wedge f/\neg v)\}$. *Then,* $P(l \wedge f/\neg v) \vee P(l \wedge \neg f/\neg v)$ *completely explains* $P(l/\neg v)$.

---

[3] The alethic modality can be eliminated in **F** and **G** (see [32]), as it reduces to $\bigcirc(\bot/\neg A)$. However, here we retain it to distinguish facts from norms.

[4] We explicitly state the permissions for $l \wedge \neg f$, $\neg l \wedge f$, and $l \wedge f$ which represent the norm's compliance conditions. As also noted by [2], some disjunctive obligations in natural language imply a permissible choice between disjuncts —a move that is not allowed in Åqvist's systems.

4 Case study: the GDPR

To see the definition at work in Åqvist systems and the issues that arise, we present the following example taken from [4]. It concerns the European Data Protection Regulation and it has the same structure as Chisholm's paradox.

Example 4.1 (GDPR) *Consider the following sentences:*

(1) *Personal data shall be processed lawfully (Art. 5),* $\bigcirc l$

(2) *If the personal data have been processed unlawfully, the controller has the obligation to erase the personal data in question without delay (Art. 17.d, right to be forgotten),* $\bigcirc(e/\neg l)$

(3) *It is obligatory, e.g., as part of a respective agreement between a customer and a company, to keep the personal data (as relevant to the agreement) provided that it is processed lawfully,* $\bigcirc(\neg e/l)$

(4) *Some data in the context of this agreement have been processed unlawfully.*

(1)-(3) give us the knowledge base $\mathcal{K}_1^{GDPR} = \{\bigcirc l, \bigcirc(e/\neg l), \bigcirc(\neg e/l)\}$. Intuitively, our logic should help explain that the data have to be erased, due to the presence of the norm (2) and the fact (4), here omitted from $\mathcal{K}_1^{GDPR}$.

4.1 A problem in Åqvist's systems

We could in principle formalize (4) as either (4.1) $\neg l$ or (4.2) $\Box\neg l$. None of them would give us the expected result. Indeed:

(4.1) By adding $\neg l$ to the knowledge base, i.e., $\mathcal{K}_2^{GDPR} = \mathcal{K}_1^{GDPR} \cup \{\neg l\}$ and putting $\bigcirc e$ as the explanandum, we do not obtain the desired explanans $\Gamma = \bigcirc(e/\neg l) \wedge \neg l$, due to the failure of FD in Åqvist's systems.

(4.2) By adding $\Box\neg l$ to $\mathcal{K}_1^{GDPR}$, we introduce an overgeneration issue, resulting in more explananda than appropriate. Using DD and SFD, we derive $\bigcirc e$ and $\bigcirc\neg e$. In **F** and **G**, this is inconsistent, because of the D* axiom, and so (by classical logic) any formula, and hence any conditional obligation $\bigcirc(B/A)$, is derivable, and then explained. In **E**, $\mathcal{K}_1^{GDPR} \cup \{\Box\neg l\}$ is satisfiable, and hence consistent. But a deontic explosion restricted to obligations under $\top$ still arises. That is, any obligation of the form $\bigcirc B$ is derivable.[5] In particular, it then follows that $\Gamma = \bigcirc l \wedge \Box\neg l$ explains $\bigcirc e$, which should not be the case.

Since van Eck [45, p. 262], this issue is known as the "dilemma of commitment and detachment" (or DD vs. FD, see Section 2). Both principles are intuitive but clash in Chisholm's scenario, forcing a choice between them.

The above example shows that replacing FD with SFD does not help, and creates more problems than it resolves: to detach $\bigcirc e$, it becomes necessary to add $\Box\neg l$ to the knowledge base, which either renders it inconsistent (in **F** and **G**) or trivializes it (in **E**). Despite this issue, one could argue that SFD is the proper detachment for CTD scenarios, as discussed in [34, §5.1] and [31, §3, §4].

[5] It follows from the derivable formula in **E**: $\bigcirc(B/A) \wedge \bigcirc(\neg B/A) \rightarrow \bigcirc(C/A)$. Put $A := \top$.

When it is settled that $\neg l$, the obligation of l no longer applies (*ought* implies *can*), necessitating the removal of this obligation from the knowledge base. Properly implementing this solution would require "updating" the knowledge base on the fly, using tools from belief change theory [17], an undertaking that lies beyond the scope of the present work.

In the next section we propose an easy way to reconcile FD and DD without updates, ensuring the right explanations.

4.2 A possible solution

The issue stems from an ambiguity in the notion of unconditional obligation. To quote Makinson [24], "unconditional obligation is deceptively ambiguous". The statement "B is obligatory" can be understood in three ways. The minimal reading, which has prevailed in deontic logic, interprets it as holding under *zero* information about the world. The minimal notion obeys DD. Thus, in Example 4.1, $\bigcirc\neg e$ tells us that $\neg e$ is the obligation the agent has under zero information about the world, in particular if it is not believed the obligation of l is violated. An alternative maximal reading—put forth by von Wright [46] and implemented in Åqvist's systems by Alchourrón [1]—would take it to mean that B is obligatory under complete information about the world. Such an interpretation can be set aside on the grounds that requiring complete information is too demanding. The intermediate reading, which seems more realistic, interprets "B is obligatory" as meaning that B is obligatory given the current information about the world. That is, it is expressed as $\bigcirc(B/A)$, where A is a Boolean formula representing a known (or believed) fact about the world. Under this intermediate notion, FD is inherently built-in.

Definition 4.2 captures this intermediary sense of "ought", which we refer to as the "detached ought". This concept reflects obligations that hold unconditionally, independent of any particular circumstances. In our notation, the operator $\bigcirc_A B$ represents such an unconditional obligation for B, where the subscript A keeps track of the fact from which the obligation is detached. It is important to emphasize that $\bigcirc_A B$ represents a genuine unconditional obligation, one that is not made relative to any conditions. In principle, for a deliberative obligation for B, it is enough that there is some A that detaches that obligation. However, the subscript also keeps track of the source of the obligation. [6]

Definition 4.2 $\mathfrak{M}, w \models \bigcirc^*_A B$ iff (i) $w \models \bigcirc(B/A)$, (ii) $w \models A$, (iii) $w \models \Diamond(A \wedge B)$, (iv) $w \models \Diamond(A \wedge \neg B)$, and (v) $w \models \Diamond\neg A$.

The following equivalence is trivially valid in **E**, **F**, and **G**:

$$\bigcirc (B/A) \wedge A \wedge \Diamond(A \wedge B) \wedge \Diamond(A \wedge \neg B) \wedge \Diamond\neg A \leftrightarrow \bigcirc^*_A B \qquad \text{(DetO)}$$

Condition (i) requires the truth of the corresponding conditional obligation, while (ii) ("actuality") the truth of the antecedent A. Conditions (iii) and (iv)

[6] This definition is a variation of a definition originally proposed by Åqvist for Chisholm's concept of "prima facie it ought to be that," utilizing propositional quantifiers [3].

('compliability" and "violability") tell us that the conditional obligation can be fulfilled or violated. (iv) blocks the validity of the implication $A \to \bigcirc A$, due to the presence of (Id). Combined with (ii), Condition (v) tells us the antecedent expresses a contingent truth−one that is true, but could have been false. Condition (v) blocks the collapse between $\bigcirc(A/\top)$ and $\bigcirc^*_\top A$. This may be justified intuitively as follows: the $\bigcirc(A/\top)$ is short for $\bigcirc A$, and does not need to be detached.

Remark 4.3 Our operator share the general idea of Ought-Implies-Can that is common in logics which explicitly link obligations to the choices available to the agent, e.g., Deontic **STIT** logic [21] and the discussion in [43]. Related (albeit different) proposals re detachment may be found in [5,28,38]. Our notion of detached obligation (and also of deliberative obligation, which will be presented in Section 5) share most similarities to the notion of instrumental obligation from [38]. [7]

With Definition 4.2 , we can obtain the expected explanations in Example 4.1, which we now demonstrate:

Example 4.4 We extend the knowledge base $\mathcal{K}_2^{GDPR}$ from Ex. 4.1 with the relevant $\Diamond$-formulas, i.e., $\mathcal{K}_3^{GDPR} = \mathcal{K}_2^{GDPR} \cup \{\Diamond\neg l, \Diamond l, \Diamond(l \wedge \neg e), \Diamond(l \wedge e), \Diamond(\neg l \wedge e), \Diamond(\neg l \wedge \neg e)\}$. [8] We obtain the desired outcome that $\Gamma = \bigcirc(e/\neg l) \wedge \neg l \wedge \Diamond l \wedge \Diamond(\neg l \wedge e) \wedge \Diamond(\neg l \wedge \neg e)$ explains $\bigcirc^*_{\neg l} e$. Also, we still have that $\Gamma = \bigcirc(\neg e/l) \wedge \bigcirc l$ explains $\bigcirc\neg e$. Hence, the conflict between FD and DD is resolved by distinguishing two sorts of "ought".

Our obligation operator also has a counterpart in terms of permission, which is not its dual and is defined as follows:

Definition 4.5 $\mathfrak{M}, w \models P^*_A B$ iff $w \models P(B/A)$, $w \models A$, $w \models \Diamond(A \wedge B)$, $w \models \Diamond(A \wedge \neg B)$, and $w \models \Diamond\neg A$.

As before, $P(B/A) \wedge A \wedge \Diamond(A \wedge B) \wedge \Diamond(A \wedge \neg B) \wedge \Diamond\neg A \leftrightarrow P^*_A B \quad (DetP)$.

Remark 4.6 In our framework, the main purpose of the alethic modality $\Box$ is not to detach obligations but rather to impose conditions on the model. For instance, it can encode meaning postulates, such as, to take a well-known example, that killing gently implies killing. In addition, it can also represent facts that cannot be changed by the agent.

5 Deliberative obligations and permissions

The primary aim of this section is to address and correct two flaws in the previous constructions, each at a different level:

[7] According to [38], instrumental obligations tell us "what we should and can bring about modulo obligations which are, concerning the factual premises, violated already, i.e., without considering already violated obligations. They tell us what is the right thing to do in a certain situation, but not what would have been the right thing to do in first place, for instance, to avoid a sub-ideal situation in which some obligations are violated."

[8] $\Diamond\neg l$ and $\Diamond l$ are not strictly necessary, since they already follow from the knowledge base, but we list them to obtain the explanation.

Technical: The notion of detached norm cannot handle (i) a norm under condition $\top$ and (ii) a specificity structure in which a more specific norm takes precedence over a more general one.

Conceptual: the construction does not take into account the difference between the context of deliberation and the context of judgment (or evaluation) [42,23]. In the former, one puts aside the moral or legal status of the facts which are taken as settled and one must decide what to do. In the context of judgment, one assesses the status of settled facts through backward looking or *post-eventum* judgments [19, p. 157].

For the purpose of explanation, we focus on the notions of deliberative obligation and permission.

Definition 5.1 [Deliberative obligation] Given a set of sentences $\mathcal{C}$ called the context, A is deliberatively obligatory iff

(i) both $\mathcal{C} \not\rhd A$ and $\mathcal{C} \not\rhd \neg A$ and either

(iiA) $\mathcal{C} \rhd \bigcirc A$, and there is no B s.t. $\mathcal{C} \rhd B$ and $\mathcal{C} \rhd P^*_B \neg A$ or

(iiB) there is a B s.t. $\mathcal{C} \rhd B$, $\mathcal{C} \rhd \bigcirc^*_B A$, and there is no D s.t. $\mathcal{C} \rhd D$, $D \rhd B$, and $\mathcal{C} \rhd P^*_D \neg A$

Condition (i) requires that A is not already known to be true or false. Condition (iiA) ensures that obligations under $\top$ are preserved, unless a more specific permission to the contrary is present. Condition (iiB) requires that $\mathcal{C}$ proves both B and $\bigcirc^*_B A$, while also ensuring that this obligation is not overridden by a more specific permission to the contrary. If the overriding norm is an obligation, in **F** and **G** $\bigcirc^*_B A \wedge \Diamond A \rightarrow P^*_D A$, so conditions (iiA) and (iiB) suffice. In **E**, this inference fails, requiring the permission to be manually added−or (iiA) and (iiB) modified to allow for either a permission or an obligation.

Remark 5.2 Definition 5.1 accounts for cases where there is a cascade of defeating obligations/permissions. For instance, for $\mathcal{K} = \{\bigcirc(B/A), \bigcirc(\neg B/A \wedge C), \bigcirc(B/A \wedge C \wedge D), A, C, D\}$, we do obtain that B is deliberatively obligatory, even though the norm $\bigcirc(B/A)$ is overridden by $\bigcirc(\neg B/A \wedge C)$. This also means that in case of a conflict of the form $\mathcal{K} = \{\bigcirc(B/A), \bigcirc(\neg B/C), A, C\}$, both B and $\neg B$ are deliberatively obligatory. There is an analogy with the treatment of unresolved conflicts in non-monotonic logic, *cf.* [37]. In our context, the so-called credulous approach asserts that each state of affairs is obligatory, while the skeptical one states that only their disjunction is. The above example with A and C shows that our definition has some elements of the credulous approach in presence of conflicting obligations under different conditions. However, Åqvist system **E** also allows for non-contradictory obligations for B and $\neg B$ under the same condition, e.g., A. In that case, our definition adopts the skeptical approach: one is not obliged to do either.

Example 5.3 *Take a version of $\mathcal{K}^{GDPR}_3$ from Example 4.4 without any fact, i.e., $\mathcal{K}^{GDPR}_4 = \mathcal{K}^{GDPR}_3 \setminus \{\neg l\}$. l and $\neg e$ are deliberatively obligatory given*

$\mathcal{C} = \mathcal{K}_4^{GDPR}$ *[condition (iiA)]. If* $\neg l$ *is the case, i.e.,* $\mathcal{C} = \mathcal{K}_3^{GDPR}$*, only* e *is deliberatively obligatory [condition (iiB)].*

By combining the notion of deliberative obligation and Definition 3.1, we obtain a notion of minimal explanation for deliberative obligations. Just replace, in Definition 3.1, A with "*A is deliberatively obligatory*" and put $\mathcal{K} = \mathcal{C}$. Then, Definition 3.3 can be used to generate appropriate complete explanations of deliberative obligations.

Example 5.4 *Take the two versions of the knowledge base from Ex. 5.3, now with the goal of generating explanations.*

Case (i) *For* $\mathcal{K}_4^{GDPR}$*, i.e., without any decided fact,* $\Gamma = \bigcirc l$ *explains why* l *is deliberatively obligatory and* $\Gamma = \bigcirc l \wedge \bigcirc(\neg e/l)$ *explains why* $\neg e$ *is deliberatively obligatory.*

Case (ii) *For* $\mathcal{K}_3^{GDPR}$*, with* $\neg l$ *included,* $\Gamma = \bigcirc(e/\neg l) \wedge \neg l \wedge \Diamond l \wedge \Diamond(\neg l \wedge e) \wedge \Diamond(\neg l \wedge \neg e)$ *explains why* e *is deliberatively obligatory.*

Turning to deliberative permission, a distinction between different notions of permission is necessary (see, e.g., [18] for an overview), in particular weak and strong permissions. The former are in place when there are no obligations to the contrary that apply ([15] discusses the complexities of this notion), while for the latter, a permission must be either explicitly present or derivable. In our explanatory setting, strong permission is more interesting, as it allows us to identify a subset of the knowledge base that explains why a particular permission is active—rather than merely noting the absence of obligations, as in the weak case. The deliberative version of strong permission is defined as:

Definition 5.5 [Deliberative strong permission] Given a set of sentences $\mathcal{C}$ called the context, A is (deliberatively) strongly permitted iff

(i) both $\mathcal{C} \not\sim A$ and $\mathcal{C} \not\sim \neg A$ and either

(ii) either $\mathcal{C} \sim PA$ and there is no B, s.t. $\mathcal{C} \sim B$ and $\mathcal{C} \sim \bigcirc_B^* \neg A$ or

(iii) there is a B, s.t. $\mathcal{C} \sim B$, $\mathcal{C} \sim P_B^* A$, and there is no D, s.t. $\mathcal{C} \sim D$, $D \sim B$, and $\mathcal{C} \sim \bigcirc_D^* \neg A$

The parallel with Definition 5.1 is that we demand that the value of A is not yet known. Then, A is strongly permitted either when it is unconditionally permitted and there is no condition in the context that makes the opposite actually obligatory, or we can obtain a detached permission for A and there is no more specific obligation to the contrary that can be detached, too.

The notion of strong permission can then be applied to our Definition 3.3 to obtain explanations of deliberative strong permissions. We replace A with *A is deliberatively strongly permitted* and take $\mathcal{K} = \mathcal{C}$.

The following example provides an application of Definition 5.5.

Example 5.6 (Motor vehicle law 3) *Consider the norms below, adapted and simplified from [40], §4 Abs. 6 Z 3, 6a.*

(1) *Motor vehicles of length above 12,00 m are forbidden to drive,*

$\bigcirc(\neg d/(>12m))$

(2) *Motor vehicles of length above 12,00 m but exceeding the limit by not more than 500 mm because of aerodynamic devices are allowed to drive,* $P(d/(>12m) \wedge (\leq 500mm) \wedge a)$

The knowledge base is $\mathcal{K}_1^{V3} = \{\bigcirc(\neg d/(>12m)), P(d/(>12m) \wedge (\leq 500mm) \wedge a)\}$, *together with the* $\Diamond$*-formulas that we omit for simplicity. The explanations are obtained as follows:*

i *In* $\mathcal{K}_1^{V3}$ *or* $\mathcal{K}_1^{V3} \cup \{(\leq 12m)\}$, *we do not obtain any deliberative obligation or permission; in particular, we do not establish that* d *is deliberatively strongly permitted.*

ii *However, arguably, a permission to drive when the vehicle's length is below or equal to 12,00 m can also be included in the formalization of the example, so that we have* $\mathcal{K}_2^{V3} = \mathcal{K}_1^{V3} \cup \{P(d/(\leq 12m))\}$.

Case (iiA) *With the vehicle's length being compliant, i.e.,* $\mathcal{K}_2^{V3} \cup \{(\leq 12m)\}$, *and using Def. 5.5 and 3.3, we have that* $P(d/(\leq 12m)) \wedge (\leq 12m)$ *indeed explains the deliberative strong permission of* d.

Case (iiB) *If the exceptional circumstance is in effect, i.e.,* $\mathcal{K}_2^{V3} \cup \{(>12m), (\leq 500mm), a)\}$, *we obtain that* $P(d/(> 12m) \wedge (\leq 500mm) \wedge a) \wedge (\leq 12m) \wedge (\leq 500mm) \wedge a$ *explains why* d *is strongly permitted. As desirable, we do not obtain an explanation of why* $\neg d$ *would be deliberatively obligatory.*

Remark 5.7 Our definitions of deliberative obligation and permission demonstrate certain non-monotonic characteristics, particularly in cases involving specificity and CTD scenarios. For instance, in Ex. 5.6, adding $(>12m)$ renders d prohibited. However, when the exceptional circumstance is also present– namely, by including $(\leq 500mm)$ and a–the prohibition is lifted. Similarly, non-monotonic behavior arises in CTD scenarios, as illustrated in Ex. 5.3.

In all the examples observed so far, the explanations for the logics **F** and **G** were identical. However, the following example demonstrates the distinct behavior of these logics.

Example 5.8 *Let us add to Example 3.4 a norm from [40], §99 Abs. 1a:*

> *Ignoring the regulation above, in a tunnel, only low-beam headlights are allowed,* $\bigcirc(l \wedge \neg f/t \wedge \neg v)$

We interpret it as being in a tunnel, only low-beam headlights are allowed even under bad visibility, since the norm explicitly refers to the norm from Ex. 3.4. Also, include the relevant $\Diamond$*-formulas (that we omit in our representation below) and take that the exceptional circumstance is in place, but also that your low-beam headlights do not work, i.e.,* $\mathcal{K}_2^{V2} = \mathcal{K}_1^{V2} \cup \{\bigcirc(l \wedge \neg f/t \wedge \neg v), \neg v, \neg l, t\}$. *First,* $P(\neg l/\neg v)$ *follows from* $P(\neg l \wedge f/\neg v)$. *Then, in* $\boldsymbol{G}$ *– but not in* $\boldsymbol{F}$*– given* (RM), *we obtain* $\bigcirc(l \vee f/\neg v \wedge \neg l)$. *Furthermore, we get the deliberative obligation of* $l \vee f$ *with the explanation* $\bigcirc(l \vee f/\neg v) \wedge P(\neg l \wedge f/\neg v) \wedge \neg v \wedge \neg l$. *However, we also obtain the deliberative obligation of* $\neg f$ *because* $\bigcirc(l \wedge \neg f/t \wedge \neg v) \wedge t \wedge \neg v$.

This shows that $\boldsymbol{G}$ *represents the example as desired: due to bad visibility, the driver is indeed obliged to* $l \vee f$*, but also to* $\neg f$*, given that they are in the tunnel. But because of* $\neg l$*, they are in a situation of deontic conflict: both norms cannot be complied with.*

6 Contrastive deontic explanations

In many cases, explanations are sought not just for why something holds but why it holds instead of an alternative. These contrastive explanations address "Why A rather than B?" questions, where A and B represent competing alternatives. In a deontic context, this contrast often involves obligations and permissions. We explore three key cases. First, one might ask why one is permitted to do (or achieve) A rather than obliged to perform (or achieve) B (when A and B are not possible together). Second, the reverse question arises: why is one obliged to do B rather than at least permitted to do A? Finally, given two obligations, $\bigcirc(B/A)$ and $\bigcirc(D/C)$, one might inquire why one is, deliberatively speaking, under the obligation to do B rather than under the obligation to do D. Our definition below focuses on facts as explananda.

Definition 6.1 [Contrastive deontic explanation] Given a knowledge base $\mathcal{K}$, fact C rather than fact D explains why one is obliged (*resp.* (strongly) permitted) to do A rather than (strongly) permitted (*resp.* obliged) to do B iff:

(i) $\mathcal{K} \rhd \bigcirc^*_C A$ (*resp.* $P^*_C A$) and

(ii) $\mathcal{K} \rhd \neg\Diamond(A \wedge B)$ and

(iii) $\mathcal{K} \rhd P(B/D)$ (resp. $\mathcal{K} \rhd \bigcirc(B/D)$) and either $\mathcal{K} \not\rhd D$ or $C \rhd D$.

The definition covers the cases mentioned above. It outputs a difference-making pair of facts that are in a counterfactual relationship: one triggers certain norms, while the other leads to different norms. In **E** we face the same problem as for point (iii) in Def. 5.1: if we demand a permission to be derived, we have to either manually add it for each obligation (given that D* does not hold) or change point (iii) in Def. 6.1.

Remark 6.2 A related notion of contrastive deontic explanation is in [44], covering the two key cases: CTDs and specificity. Indeed, (iii) tells us that, if the KB entails e.g., an obligation $\bigcirc(B/D)$ potentially overriding the permission, then either this obligation is not triggered ($\mathcal{K} \not\rhd D$) or it is more general ($C \rhd D$). This will be the case in case (ii) of Example 6.3 below. We do not consider the case of a contrast between two permissions, however. This is because one can be permitted to do both A and B, even if they are incompatible.

We now illustrate two contrastive cases with an example.

Example 6.3 *Consider the case presented in Example 5.6.*

Case (i) $\mathcal{K}_1^{V3} \cup \{(>12m)\}$. $(>12m)$ *rather than* $(>12m) \wedge (\leq 500mm) \wedge a$ *explains why one is obliged to* $\neg d$ *rather than strongly permitted to* d*.*

Case (ii) $\mathcal{K}_1^{V3} \cup \{(>12m), (\leq 500mm), a\}$. $(>12m) \wedge (\leq 500mm) \wedge a$ *rather than* $(>12m)$ *explains why* d *is strongly permitted rather than forbidden.*

Remark 6.4 Also contrastive explanations have some non-monotonicity features, *cf.* Rem. 5.7. This can be seen by comparing cases (i) and (ii) in Ex. 6.3: adding a fact results in the retraction of a conclusion about an explanandum.

7 Norms and normality

To demonstrate the flexibility of our framework, we present an example involving a generalized quantifier "most." To handle these, we adopt the extension of Åqvist's systems from [29], incorporating a normality conditional $\Rightarrow$. The expression $A \Rightarrow B$ is interpreted as "If A, then normally B". Explanations frequently arise from the interplay between normative and normality statements. To achieve this, we adapt our framework to work with the semantic construction in [29] (see Appendix 8 for a summary). We illustrate the construction with an example, which incorporates a specificity structure of a kind that that cannot be accounted in Åqvists's systems, and most existing accounts. Note that normality conditionals count as facts in our framework, given that they are non-deontic formulas. That way, explanations can now also include statements about normality.

The example below adapts and simplifies a case in the Court of Justice of the European Union's ruling [39], involving a dispute between Germany and Austria. Austria argued that Germany violated the EU obligation to treat all European citizens equally by requiring vignettes while simultaneously granting a tax benefit of the same amount exclusively for owners of vehicles registered in Germany—who are predominantly German. In our formalization, we primarily reference Sect. 3.3 and §38-42 and §47-52 of the judgment, focusing solely on "the first ground of complaint" outlined in the ruling.

Example 7.1 (Highway) *Consider the following sentences, where t, u, r, g stand for: paying a toll, using the highway, having a car registered in Germany, and being of German nationality, respectively:*

(i) *Users of highways should pay a toll:* $\bigcirc(t/u)$;

(ii) *Users of highways having their car registered in Germany may not pay a toll:* $P(\neg t/u \wedge r)$;

(iii) *Most people having their car registered in Germany are German,* $r \Rightarrow g$;

(iv) *Most people not having their car registered in Germany are non-German,* $\neg r \Rightarrow \neg g$.

In the (extended) system, we will derive, as the court does:

(v) *Normally, only non-German (users) are obliged to pay a toll, while German (users) are permitted not to:* $\bigcirc(t/u \wedge \neg g) \wedge P(\neg t/u \wedge g)$.

This represents, the court argues, indirect discrimination based on nationality.

We first describe, using the example, how the logic defined in [29] works. In modal logic, the (local) semantic consequence relation preserves truth within

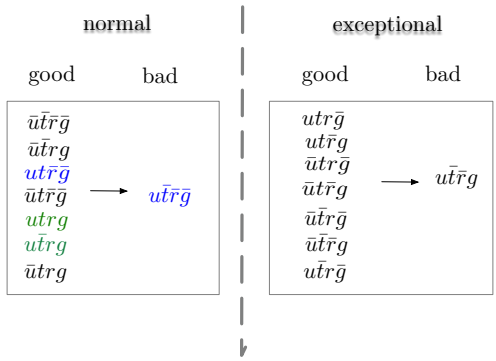

Fig. 1. Model $\mathcal{M}$ "induced" by (i)-(iv). Colors indicate the worlds relevant for the truth of (v). A propositional letter with a bar above it represents its negation.

a specific world. However, not just any model suffices—the ranking of the required model is further constrained by the conditionals in the premises, namely (i)-(iv). Figure 1 shows this model, we call it $\mathcal{M}$. First, the possible worlds have been ranked based on how well they comply with the normality conditionals (iii) and (iv). The boxes denote the resulting equivalence classes in the normality ordering. The box on the left shows the "normal" or typical worlds, while the box on the right shows the "exceptional" or atypical worlds. In these exceptional worlds, for instance, German natives do not have their car registered in Germany, because they do not live in their home country. These worlds are exceptional, because most of the German natives do live in their home country, and so they have their car registered there. Within each cluster, the worlds have been ordered based how well they comply with the obligations (i) and (ii).

Note that in itself the ideality ranking is independent of the normality ranking. The normality dimension plays a role only when evaluating an obligation. In models of this sort, roughly speaking $\bigcirc(B/A)$ (resp. $P(B/A)$) holds true, whenever the most normal $A \wedge B$-world is strictly better than (resp. at least as good as) the most normal $A \wedge \neg B$-world. One can see that $\bigcirc(t/u \wedge \neg g)$, because the most normal $u \wedge \neg g \wedge t$ is strictly better than the most normal $u \wedge \neg g \wedge \neg t$ (they are both highlighted in blue). We also have $P(\neg t/u \wedge g)$, because the most normal $u \wedge g \wedge \neg t$ is at least as good as the most normal $u \wedge g \wedge t$ (they are both highlighted in green). Exceptional worlds that satisfy such conjunctions are, in a crucial way, disregarded.

Let $\mathcal{K}_1^H$ be $\{\bigcirc(t/u), P(\neg t/u \wedge r), r \Rightarrow g, \neg r \Rightarrow \neg g, \Diamond\neg u, \Diamond(u \wedge r \wedge \neg t), \Diamond(u \wedge r \wedge t)\}$. Observe that $\mathcal{K}_1^H \rhd \bigcirc(t/u \wedge \neg g) \wedge P(\neg t/u \wedge g)$, i.e. (v), where here and below $\rhd$ stands for the consequence relation $\Vdash$ from [29] (see Appendix 8).

We can now provide contrastive explanations that align with the court's reasoning (for conciseness, we omit the adjectives "deliberatively" and "strongly"):

(1) $\mathcal{K}_1^H \cup \{u, g\}$, i.e., German users are considered. In this case $u \wedge g$ rather than $u \wedge \neg g$ explains why $\neg t$ is permitted instead of t obligatory.

(2) $\mathcal{K}_1^H \cup \{u, \neg g\}$, i.e., non-German users are considered. Now $u \wedge \neg g$ rather than $u \wedge g$ explains why t is obligatory rather than $\neg t$ permitted.

To see the unique aspect of the specificity structure presented in this scenario, it is essential to look at it from a rule-based perspective, where one builds the extension by a (possibly controlled) step-by-step application of triggered rules.[9] It is striking that in case (1) the antecedent of (ii) is not (classically) entailed by the facts in $\mathcal{K}_3^H$, and so strictly speaking the overriding norm (ii) is not "triggered" by the input.

From the court's perspective, these two explanations together form the basis for indirect discrimination: although norms (i) and (ii) do not explicitly target nationality, their effect is similar—being (or not being) German influences whether a particular norm applies.

8 Conclusion and future work

This paper presented a framework for deontic explanations that effectively handles violations and exceptions, and also provides the flexibility to accommodate various deontic logics. It has been implemented using Åqvist's dyadic deontic systems **E**, **F**, and **G**, and validated through several legal examples. Some limitations of the original systems were identified, leading to extending them. Future work includes:

(i) to investigate the logical properties of the newly introduced operators

(ii) to test the framework on a wider set of examples

(iii) to compare the framework on different logics

(iv) to investigate its suitability for counterfactual explanations

(v) to investigate the possibility of reintroducing SFD, and compare the two approaches

Ad (iii). The framework has fruitfully been used to show the differences in explanatory power between **E**, **F**, and **G**. We would like to investigate if it can be used to differentiate Åqvist's systems and other mainstream systems, like the input/output logic of [26].

Ad (iv). Since Pearl [33], counterfactual explanations (identifying input changes that alter outcomes) have become a common approach to explaining decisions made by AI systems. An example of a counterfactual explanation is: If I had a permanent job, my loan would have been approved. Sometimes, the focus shifts from factual changes to required obligations. Our notion of contrastive explanation already includes a counterfactual element, warranting further study on its link to AI explanations.

Ad (v). Many see SFD as the proper detachment principle for CTD scenarios. The idea is to make the knowledge base updatable on the fly, using tools from belief revision.

[9] This is the standard approach in the area of nonmonotonic reasoning, called the "increment" approach in [25,30].

References

[1] Alchourrón, C., *Philosophical foundations of deontic logic and the logic of defeasible conditionals*, in: J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*, John Wiley & Sons, Inc., New York, 1993 pp. 43–84.

[2] Aloni, M., *Free choice, modals, and imperatives*, Natural Language Semantics **15** (2007), pp. 65–94.

[3] Åqvist, L., "Prima facie obligations in deontic logic: A Chisholmian analysis based on normative preference structures," De Gruyter, Berlin, New York, 1998 pp. 135–155.

[4] Benzmüller, C., X. Parent and L. W. N. van der Torre, *A deontic logic reasoning infrastructure*, in: F. Manea, R. G. Miller and D. Nowotka, editors, *Sailing Routes in the World of Computation - 14th Conference on Computability in Europe, CiE 2018*, Lecture Notes in Computer Science **10936** (2018), pp. 60–69.

[5] Carmo, J. and A. J. I. Jones, *Deontic logic and contrary-to-duties*, in: D. M. Gabbay and F. Guenthne, editors, *Handbook of Philosophical Logic, 2nd ed. (Vol.8)*, Springer, 2002 pp. 265—343.

[6] Chellas, B., "Modal Logic," Cambridge University Press, Cambridge, 1980.

[7] Chisholm, R., *Contrary-to-duty imperatives and deontic logic*, Analysis **24** (1963), pp. 33–63.

[8] Ciabattoni, A., N. Olivetti and X. Parent, *Dyadic obligations: proofs and countermodels via hypersequents* (2022), pp. 54–71.

[9] Ciabattoni, A., N. Olivetti, X. Parent, R. Ramanayake and D. Rozplokhas, *Analytic Proof Theory for Åqvist's System F*, in: C. P. Juliano Maranhão, C. Straßer and L. van der Torre, editors, *Deontic Logic and Normative Systems - 16th International Conference, DEON 2023* (2023), pp. 79–98.

[10] Ciabattoni, A. and M. Tesi, *Sequents vs hypersequents for Åqvist systems*, in: C. Benzmüller, M. J. H. Heule and R. A. Schmidt, editors, *IJCAR 2024*, 2024, p. 176–195.

[11] Davidson, D., "Essays on Actions and Events," OUP, 1980.

[12] Florio, C. D., A. Rotolo, G. Governatori and G. Sartor, *Stable normative explanations: From argumentation to deontic logic*, in: M. O. S. Gaggl, M.V. Martinez, editor, *Logics in Artificial Intelligence. JELIA 2023* (2023), pp. 123–131.

[13] Gabbay, D., J. Horty, X. Parent, R. van der Mayden and L. van der Torre, editors, "Handbook of Deontic Logic and Normative Systems, Volume 2," College Publications, 2021.

[14] Gabbay, D., J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, "Handbook of Deontic Logic and Normative Systems," College Publications, 2013.

[15] Governatori, G., *Weak permission is not well-founded, grounded and stable*, CoRR **abs/2411.10624** (2024).
URL `https://doi.org/10.48550/arXiv.2411.10624`

[16] Governatori, G., F. Olivieri, A. Rotolo and M. Cristani, *Stable normative explanations*, in: E. Francesconi, G. Borges and C. Sorge, editors, *Legal Knowledge and Information Systems - JURIX 2022*, IOS Press, 2022 pp. 43–52.

[17] Gärdenfors, P., "Knowledge in Flux," MIT Press, 1988.

[18] Hansson, S. O., *The varieties of permission*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems (Vol.1)*, College Publications, 2013 pp. 195–240.

[19] Hare, R., "The Language of Morals," Clarendon, 1991.

[20] Hempel, C., "Aspects of Scientific Explanation and other Essays in the Philosophy of Science," Free Press, New York, 1965.

[21] Horty, J. F., "Agency and Deontic Logic," Oxford University Press, 2001.

[22] Lehmann, D., *Another perspective on default reasoning*, Ann Math Artif Intell **15** (1995).

[23] Loewer, B. and M. Belzer, *Dyadic deontic detachment*, Synthese **54** (1983), pp. 295–318.

[24] Makinson, D., *On a fundamental problem of deontic logic*, in: P. Mc Namara and H. Prakken, editors, *Norms, Logics and Information Systems*, Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, 1999 pp. 29–53.

[25] Makinson, D., "Bridges from Classical to Nonmonotonic Logic," College Publications, London, 2005.
[26] Makinson, D. and L. W. N. van der Torre, *Input/output logics*, J. Philos. Log. **29** (2000), pp. 383–408.
[27] Miller, T., *Explanation in artificial intelligence: Insights from the social sciences*, Artificial Intelligence **267** (2019), pp. 1–38.
[28] Nute, D. and X. Yu, *Introduction*, in: D. Nute, editor, *Defeasible Deontic Logic. Synthese Library, vol 263*, Springer, 1997 pp. 1–16.
[29] Parent, X., *On a problem of J. Horty*, under review.
[30] Parent, X., *Moral particularism in the light of deontic logic*, Artificial Intelligence and Law **19** (2011), pp. 75–98.
[31] Parent, X., *Why be afraid of identity?*, in: A. Artikis, R. Craven, N. K. Cicekli, B. Sadighi and K. Stathis, editors, *Logic Programs, Norms and Action - Essays in Honor of Marek J. Sergot on the Occasion of His 60th Birthday*, Lecture Notes in Artificial Intelligence **7360** (2012), pp. 295–307.
[32] Parent, X., *Preference semantics for dyadic deontic logic: a survey of results*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems (Vol.2)*, College Publications, 2021 pp. 7–69.
[33] Pearl, J., "Causality: Models, Reasoning, and Inference," Cambridge University Press, 2009, 2nd edition.
[34] Prakken, H. and M. Sergot, *Dyadic deontic logic and contrary-to-duty obligations*, in: D. Nute, editor, *Defeasible Deontic Logic*, Kluwer, Dordrecht, 1997 pp. 223–262.
[35] Rotolo, A. and G. Sartor, *Argumentation and explanation in the law*, Frontiers in Artificial Intelligence **6** (2023).
[36] Rozplokhas, D., *LEGO-like small-model constructions for Åqvist's logics*, Proceedings of AIML 2024 (2024).
[37] Strasser, C. and G. A. Antonelli, *Non-monotonic Logic*, in: E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, 2024, Winter 2024 edition .
[38] Straßer, C., *A deontic logic framework allowing for factual detachment*, Journal of Applied Logic **9** (2011), pp. 61–80.
[39] The Court of Justice of the European Union, *Judgment of the court (grand chamber) of 18 june 2019. Republic of Austria v Federal Republic of Germany. Case C-591/17. ECLI identifier: ECLI:EU:C:2019:504* (2019), `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62017CJ0591`.
[40] The National Council of Austria, *Bundesgesetz vom 23. Juni 1967 über das Kraftfahrwesen (Kraftfahrgesetz 1967 – KFG. 1967)* (1967), `https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10011384`.
[41] The National Council of Austria, *Bundesgesetz über die Mauteinhebung auf Bundesstraßen (Bundesstraßen-Mautgesetz 2002 – BStMG)* (2002), `https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20002090`.
[42] Thomason, R. H., *Deontic logic and the role of freedom in moral deliberation*, in: R. Hilpinen, editor, *New Studies in Deontic Logic: Norms, Actions, and the Foundations of Ethics*, Wiley-Blackwell, Dordrecht, Netherlands, 1981 pp. 177–186.
[43] van Berkel, K. and T. Lyon, *The varieties of ought-implies-can and deontic stit logic*, in: F. Liu, A. Marra, P. Portner and F. V. D. Putte, editors, *Proceedings of 15th International Conference, DEON 2020/2021*, College Publications, 2021 pp. 55–76.
[44] van Berkel, K. and C. Straßer, *Towards deontic explanations through dialogue*, in: T. Kampik, K. Cyras, A. Rago and O. Cocarascu, editors, *Proceedings of the 2nd International Workshop on Argumentation for eXplainable AI*, CEUR Workshop Proceedings **3768** (2024), pp. 29–40.
[45] van Eck, J. A., *A system of temporally relative modal and deontic predicate logic and its philosophical applications*, Logique et Analyse **25** (1982), pp. 249–290.

[46] von Wright, G. H., editor, "An Essay in Deontic Logic and the General Theory of Action," North-Holland Pub. Co., Amsterdam, 1968.
[47] Åqvist, L., *Deontic logic*, in: R. Hilpinen, editor, *Handbook of Philosophical Logic, Volume II*, Springer, Dordrecht, Netherlands, 1984 pp. 605–714.

Appendix

The material below is taken from [29]. We describe it to make the paper self-contained. The generic term 'rule' is used to refer to a conditional obligation and a conditional statement of normality, interchangeably. The symbol r is utilized to denote either of these without distinction. $h(r)$ is the head (consequent) of the rule and $b(r)$ is its body (antecedent). $\mathbf{R}$ denotes a finite set of rules. This is the normative code, supplemented with normality statements. $\mathbf{R}^{\Rightarrow}$ is the subset of rules of the $\Rightarrow$-type, and $\mathbf{R}^{O}$ is the subset of rules of the $\bigcirc$-type. Any $P(B/A) \in \mathbf{R}^{O}$ is short for $\bigcirc(B/A \wedge B)$.

Definition 8.1 Given $r, r' \in \mathbf{R}^{\Rightarrow}$, and a set C of background assumptions, we say that r overrides r' (notation: $r \triangleright r'$) whenever

(i) $h(r), h(r'), C \vdash \bot$ (the heads of the rules are inconsistent)

(ii) $b(r) \vdash b(r')$ and $b(r') \nvdash b(r)$ (r is more specific than r')

For $r, r' \in \mathbf{R}^{O}$, we add the following extra condition:

(iii) $h(r), b(r') \nvdash \bot$ (r' is not a CTD of r).

The account in [29] uses Lehmann [22]'s lexicographic ordering for normality. For simplicity's sake, we use a simpler version, where the level of specificity among normality conditionals does not matter. It has no bearing on the treatment of the example.

Below, $V_{\Rightarrow}(w, \mathbf{R})$ and $V_o(w, \mathbf{R})$ gather the set of normality conditionals and deontic conditionals violated by w, respectively.

Definition 8.2 [Violation sets 1] Set

$$\begin{aligned} V_{\Rightarrow}(w, \mathbf{R}) =& \{r \in \mathbf{R}^{\Rightarrow} : w \models b(r) \wedge \neg h(r)\} \\ V_o(w, \mathbf{R}) =& \{r \in \mathbf{R}^{O} : w \models b(r) \wedge \neg h(r) \text{ and } w \not\models b(r') \\ & \forall r' \in \mathbf{R}^{O} \text{ s.t. } r' \triangleright r\} \end{aligned}$$

The basic idea of the construction is to impose more structure on the model using the conditionals in $\mathbf{R}$. An $\mathbf{R}$-ordered model has the form

$$\mathcal{M} = (W, \succeq_N, \succeq_I, v)$$

where W is a (non-empty) set of possible worlds, $\succeq_N$ and $\succeq_I$ denote a normality and ideality ordering of the possible worlds, and v a valuation function assigning to each possible world the set of atomic propositions true in this world. "$w \succeq_N v$" is read as "w is at least as normal as v", and "$w \succeq_I v$" is read as "w is at least as good (or ideal) as v". They are defined as in Def. 8.3 and 8.4, respectively. Given $X \subseteq W$, $\max_{\succeq_N}(X)$ is the set of maximal elements of X under $\succeq_N$.

Definition 8.3 [Levels of normality] For a set $\mathbf{R}$ of rules, and $w_1, w_2 \in W$, set:

$$w_1 \succeq_N w_2 \text{ iff } V_{\Rightarrow}(w_1, \mathbf{R}) \subseteq V_{\Rightarrow}(w_2, \mathbf{R}) \tag{1}$$

$$\max_1 = \max_{\succeq_N}(W) \tag{2}$$

$$\max_i = \max_{\succeq_N}(W - \bigcup_{k=1}^{i-1} \max_k) \tag{3}$$

Definition 8.4 [Ideality] For a set $\mathbf{R}$ of rules, and $w_1, w_2 \in W$, set:

$$w_1 \succeq_I w_2 \text{ iff: } V_o(w_1, \mathbf{R}) \subseteq V_o(w_2, \mathbf{R}) \tag{4}$$

Definition 8.5 $U \succeq_I^s V$ (reading: U is at least as good as V, the superscript s is for "set") whenever for all $v \in V$ there is some $u \in U$ s.t. $u \succeq_I v$. $U \succ_I^s V$ is a shorthand for $V \not\succeq_I^s U$.

Intuitively, $A \Rightarrow B$ holds in w if all the most normal A-worlds are B-worlds. And $\bigcirc(B/A)$ (resp. $P(B/A)$) holds if the set of most normal $A \wedge B$-worlds is strictly better than (resp. "at least as good as") the set of most normal $A \wedge \neg B$-worlds. Formally:

Definition 8.6 [Truth-conditions]

$$M, w \models A \Rightarrow B \text{ iff } \max_{\succeq_N}(\|A\|) \subseteq \|B\| \tag{5}$$

$$M, w \models \bigcirc(B/A) \text{ iff } \max_{\succeq_N}(\|A \wedge B\|) \succ_I^s \max_{\succeq_N}(\|A \wedge \neg B\|) \tag{6}$$

$$M, w \models P(B/A) \text{ iff } \max_{\succeq_N}(\|A \wedge B\|) \succeq_I^s \max_{\succeq_N}(\|A \wedge \neg B\|) \tag{7}$$

where $\|A\|$ is the truth-set of A, viz. the set of worlds where A holds.

The consequence relation preserves truth-in-a-world from the premises to a conclusion, but in models whose ordering is obtained as described above. (This version slightly simplifies the definition found in [29].)

Definition 8.7 [Consequence] Where $\mathbf{R} \subseteq \Gamma$, $\Gamma \Vdash A$ iff $\Gamma \models_{\mathcal{C}(\mathbf{R})} A$, where $\mathcal{C}(\mathbf{R})$ denotes the class of $\mathbf{R}$-ordered models.