Automated Verification of Deontic Correspondences in Isabelle/HOL - First Results

Xavier Parent[1], Christoph Benzmueller[2,3]

[1]Technische Universität Wien
[2]Universität Bamberg
[3]Freie Universität Berlin

Abstract
We report our first results regarding the automated verification of deontic correspondences (broadly conceived) and related matters in Isabelle/HOL, analogous to what has been achieved for the modal logic cube.

Keywords
Correspondence, betterness, dyadic deontic logic, conditional obligation, automated reasoning, Isabelle/HOL

1. Introduction

We report our first results regarding the automated verification of deontic correspondences (broadly conceived) and related matters in Isabelle/HOL, analogous to what has been achieved for the modal logic cube in [1]. To look at Standard Deontic Logic (SDL) and extensions [2, 3] would not be very interesting. First, no new insights would be gained, since SDL is a normal modal logic of type KD. Second SDL is vulnerable to the well-known deontic paradoxes, like in particular Chisholm's paradox of contrary-to-duty obligation (see [3] for details). We focus on the dyadic deontic logics with a preference-based semantics, which originate from the work of Danielsson, Hansson, van Fraassen, Lewis and others. One uses an "intensional" conditional to represent conditional obligation sentences that is weaker than the one obtained using material implication. The semantics generalizes that of SDL, by allowing for grades of ideality. That framework is particularly popular in deontic logic, see the overview chapter in the second volume of the *Handbook of Deontic Logic* [4]. In that framework, a preference relation $\succeq$ ranks the possible worlds in terms of comparative goodness or betterness.[1] The conditional obligation of ψ, given φ (notation: $\bigcirc(\psi/\varphi)$) is evaluated as true if the best φ-worlds are all ψ-worlds. Like in modal logic, different properties of the betterness relation yield different systems. So far the correspondence between properties and modal axioms have been established "with pen and paper". This raises the question of how much of this correspondence can be automated. As explained in [1] we believe that "automation facilities could be very useful for the exploration of

ARQNL 2022: Automated Reasoning in Quantified Non-Classical Logics, 11 August 2022, Haifa, Israel
xavier@logic.at (X. Parent); christoph.benzmueller@uni-bamberg.de (C. Benzmueller)
https://xavierparent.co.uk (X. Parent); http://christoph-benzmueller.de (C. Benzmueller)

[1]For $i \succeq j$, read "i is at least as good as j".

the meta-theory of other logics, for example, conditional logics, since the overall methodology is obviously transferable to other logics of interest". Here we follow up on that suggestion, building on results from [5], where the weakest available system (called **E**) has faithfully been embedded in Higher-Order Logic (HOL). In the present paper we consider extensions of **E**, already identified with pen and paper. We look at connections or correspondences between axioms and semantic conditions as "extracted" by relevant soundness and completeness theorems. Thus, we take "correspondence" in the same (broad) sense as Hughes and Cresswell, who write:

> "D, T, K4, KB [are] produced by adding a single axiom to K and [...] in each case the system turns out to be characterized by [sound and complete wrt] the class of models in which [the accessibility relation] R satisfies a certain condition. When such a situation obtains–i.e. when a system K+α is characterized by the class of all models in which R satisfies a certain condition—we shall [...] say [...] that the wff α itself is characterized by that condition, or that the condition *corresponds* [their italics] to α." [6, p. 41]

The theory file we discuss is available for downloading at http://logikey.org under sub-repository "/Deontic-Logics/cube-dll/" (file "cube.thy").

The paper is organized as follows. Section 2 recalls system **E** and its extensions. Section 3 shows the embedding of **E** in Isabelle/HOL. Section 4 describes the encoding of the properties of the betterness relation. Section 5 studies the correspondence between the latter properties and the axioms. Section 6 looks at a well-known alternative evaluation rule for the conditional put forth by Lewis [7]. Section 7 concludes.

2. System E

We describe the semantics and proof theory of system E and its extensions. This one introduces the primitive symbol $\bigcirc(\_/\_)$ for "it is obligatory that ... given that ...", from which symbol $P(\_/\_)$ for "it is permitted that ... given that ..." is defined. The language also has $\Box$ and $\Diamond$.

2.1. Semantics

We start with the main ingredients of the semantics. A preference model is a structure $M = (W, \succeq, V)$, where W is a non-empty set of possible worlds, $\succeq$ is a preference relation ranking elements of W in terms of betterness or comparative goodness, and V is a function assigning to each propositional letter a subset of W (intuitively, the subset of those worlds where the propositional letter is true). $a \succeq b$ may be read "a is at least as good as b". $\succ$ is the strict counterpart of $\succeq$, defined by $a \succ b$ (a is strictly better than b) iff $a \succeq b$ and $b \not\succeq a$. $\approx$ is the equal goodness relation, defined by $a \approx b$ (a and b are equally good) iff $a \succeq b$ and $b \succeq a$.

The truth conditions for modal and deontic formulas read:

- $M, a \vDash \Box\varphi$ iff $\forall b \in W$ we have $M, b \vDash \varphi$
- $M, a \vDash \bigcirc(\psi/\varphi)$ iff $\forall b \in \mathrm{best}(\varphi)$ we have $M, b \vDash \psi$

When no confusion can arise, we omit the reference to M and simply write $a \models \varphi$. Intuitively, $\bigcirc(\psi/\varphi)$ is true if the best φ-worlds are all ψ-worlds. There is variation among authors regarding the formal definition of "best". It is sometimes cast in terms of maximality (we call this the max rule) and some other times cast in terms of optimality (we call this the opt rule). An φ-world a is maximal if it is not (strictly) worse than any other φ-world. It is optimal if it is at least as good as any φ-world. The two notions coincide only when "gaps" (incomparabilities) in the ranking are ruled out. Formally:

Max rule	**Opt rule**
$\mathrm{best}(\varphi) = \max(\varphi)$	$\mathrm{best}(\varphi) = \mathrm{opt}(\varphi)$

where

$$a \in \max(\varphi) \Leftrightarrow a \models \varphi \;\&\; \neg\exists b\, (b \vDash \varphi \;\&\; b \succ a)$$
$$a \in \mathrm{opt}(\varphi) \Leftrightarrow a \models \varphi \;\&\; \forall b\, (b \vDash \varphi \rightarrow a \succeq b)$$

The relevant properties of $\succeq$ are (universal quantification over worlds is left implicit):

- Reflexivity: $a \succeq a$;
- Transitivity: if $a \succeq b$ and $b \succeq c$, then $a \succeq c$;
- Totalness or (strong) connectedness: $a \succeq b$ or $b \succeq a$ (or both);
- Interval order: $\succeq$ is reflexive and Ferrers (if $a \succeq b$ and $c \succeq d$, then $a \succeq d$ or $c \succeq b$).

The interval order condition makes room for the idea of non-transitive equal goodness relation due to discrimination thresholds. These are cases where $a \approx b$ and $b \approx c$ but $a \not\approx c$ (see [8]).

Lewis' limit assumption is meant to rule out sets of worlds without a "limit" (viz. a best element). Its exact formulation varies among authors. It exists in (at least) the following four versions, where $\mathrm{best} \in \{\max, \mathrm{opt}\}$

Limitedness
If $\exists x$ s.t. $x \models \varphi$ then $\mathrm{best}(\varphi) \neq \emptyset$ (LIM)

Smoothness (or stopperedness)
If $x \models \varphi$, then: either $x \in \mathrm{best}(\varphi)$ or $\exists y$ s.t. $y \succ x \;\&\; y \in \mathrm{best}(\varphi)$ (SM)

A betterness relation $\succeq$ will be called "opt-limited" or "max-limited" depending on whether (LIM) holds with respect to opt or $\max$. Similarly, it will be called "opt-smooth" or "max-smooth" depending on whether (SM) holds with respect to opt or $\max$. For pointers to literature, and the relationships between these versions of the limit assumption, see [9, 4].

The above semantics may be viewed as a special case of the selection function semantics flavored by Stalnaker and generalized by Chellas [10]. The preference relation is replaced with a selection function f from formulas to subsets of W, such that, for all φ, $f(\varphi) \subseteq W$. Intuitively, $f(\varphi)$ outputs all the best φ-worlds. The evaluation rule for the dyadic obligation operator is phrased thus: $\bigcirc(\psi/\varphi)$ holds when $f(\varphi) \subseteq \|\psi\|$, where $\|\psi\|$ is the set of ψ-worlds. It is known that when suitable constraints are put on the selection function, the two semantics

validate exactly the same set of formulas. See [11, 4] for details.[2] The correspondences between constraints put on the selection function and modal axioms are verified by automated means in [12]. A comparison between the two studies is left as a topic for future research.

2.2. Systems

The relevant systems are shown in Fig. 1. A line between two systems indicates that the system to the left is strictly included in the system to the right.

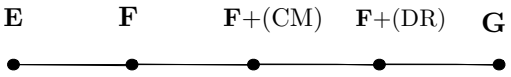

Figure 1: Systems

All contain the classical propositional calculus; they then add the following schemata:

- For **E** (the naming follows [4]):

S5-schemata for $\Box$	(S5)
$\bigcirc(\psi \to \xi/\varphi) \to (\bigcirc(\psi/\varphi) \to \bigcirc(\xi/\varphi)$	(COK)
$\bigcirc(\psi/\varphi) \to \Box\bigcirc(\psi/\varphi)$	(Abs)
$\Box\varphi \to \bigcirc(\varphi/\psi)$	(Nec)
$\Box(\varphi \leftrightarrow \psi) \to (\bigcirc(\xi/\varphi) \leftrightarrow \bigcirc(\xi/\psi))$	(Ext)
$\bigcirc(\varphi/\varphi)$	(Id)
$\bigcirc(\xi/\varphi \wedge \psi) \to \bigcirc(\psi \to \xi/\varphi)$	(Sh)
If $\vdash \varphi$ then $\vdash \Box\varphi$	(N)

- For **F**: axioms of **E** plus

$$\Diamond\varphi \to (\bigcirc(\psi/\varphi) \to P(\psi/\varphi)) \qquad (D^\star)$$

- For **F**+(CM): axioms of **F** plus

$$(\bigcirc(\psi/\varphi) \wedge \bigcirc(\xi/\varphi)) \to \bigcirc(\xi/\varphi \wedge \psi) \qquad (CM)$$

- For **F**+(DR): axioms of **F** plus

$$\bigcirc(\xi/\varphi \vee \psi) \to (\bigcirc(\xi/\varphi) \vee \bigcirc(\xi/\psi)) \qquad (DR)$$

- For **G**: axioms of **F** plus:

$$(P(\psi/\varphi) \wedge \bigcirc(\psi \to \xi/\varphi)) \to \bigcirc(\xi/\varphi \wedge \psi) \qquad (Sp)$$

We give an intuitive explanation for these axioms. (COK) is the conditional analogue of the familiar distribution axiom K. (Abs) is the absoluteness axiom of [7], and reflects the fact that the ranking is not world-relative. (Nec) is the deontic counterpart of the familiar necessitation rule.

[2] One can go one step further, and make the selection function semantics an instance of a more general semantics equipped with a neighborhood function, like in traditional modal logic (see [10, §8]).

(Ext) permits the replacement of necessarily equivalent formulas in the antecedent of deontic conditionals. (Id) is the deontic analogue of the identity principle. (D*) rules out the possibility of conflicts between obligations, for a "consistent" context A. (CM) and (DR) correspond to the principle of cautious monotony and disjunctive rationality from the non-monotonic logic literature. (CM) tells us that complying with an obligation does not modify the other obligations arising in the same context. (DR) tells us that if a disjunction of states of affairs triggers an obligation, then at least one disjunct triggers this obligation. Due to Spohn, (Sp) is equivalent with the principle of rational monotony; $\bigcirc(\psi \to \xi/\varphi)$ is changed into $\bigcirc(\xi/\varphi)$. The principle says that realizing a permission does not modify the other obligations arising in the same context.

For more background on these systems, see [4] and the references therein.

2.3. Correspondences

Table 1 shows some of the known "correspondences" between semantic properties and formulas. The leftmost column shows the properties of $\succeq$. The two middle columns show the corresponding modal axioms, the first column for when the max rule is used, and the second one for when the opt rule is used. It is understood that smoothness (resp. limitedness) is defined for max in the max column, and for opt in the opt column. The rightmost column gives the paper where the completeness theorem is established. The symbol $\times$ indicates that the property (or pair of properties) is known not to correspond to any axiom. On the fifth line the parenthesis "(+smoothness)" indicates that smoothness is assumed in the background.[3]

Property	Formula (max)	Formula (opt)	Reference
reflexivity	$\times$	$\times$	[11]
totalness	$\times$	$\times$	[11]
limitedness	D*	D*	[11]
smoothness	CM	CM	[9]
transitivity (+smoothness)	$\times$	Sp	[13, 9]
transitivity+totalness	Sp	$\times$	[9]
interval order	DR	DR	[13]

Table 1
Some correspondences

3. System E in Isabelle/HOL

Our modelling of System **E** in Isabelle/HOL reuses and adapts prior work [5] and it instantiates and applies the LogiKEy methodology [14], which supports plurality at different modelling layers.

[3] Even though smoothness does not play any apparent role in the validation of the axiom, the completeness result is for a class of models satisfying this property.

3.1. LogiKEy

Classical higher-order logic (HOL) is fixed in the LogiKEy methodology and infrastructure [14] as a *universal meta-logic* [15] at the base layer (L0), on top of which a plurality of (combinations of) object logics can become encoded (layer L1). In the case of this paper, we encode extensions of System **E** at layer L1 in order to assess them. Employing these object logics notions of layer L1 we can then articulate a variety of logic-based domain-specific languages, theories and ontologies at the next layer (L2), thus enabling the modelling and automated assessment of different application scenarios (layer L3). Note that the assessment studies conducted in this paper at layer L3 do not require any further knowledge to be provided at layer L2; hence layer L2 modellings do not play a role in this paper.

LogiKEy significantly benefits from the availability of theorem provers for HOL, such as Isabelle/HOL which internally provides powerful automated reasoning tools such *Sledgehammer* [16, 17] and *Nitpick* [18]. The automated theorem proving systems integrated via *Sledgehammer* include higher-order ATP systems, first-order ATP systems, and SMT (satisfiability modulo theories) solvers, and many of these systems in turn use efficient SAT solver technology internally. Indeed, proof automation with *Sledgehammer* and (counter)model finding with *Nitpick* were invaluable in supporting our exploratory modeling approach at various levels. These tools were very responsive in automatically proving (*Sledgehammer*), disproving (*Nitpick*), or showing consistency by providing a model (*Nitpick*). In the first case, references to the required axioms and lemmas were returned (which can be seen as a kind of abduction), and in the case of models and counter-models they often proved to be very readable and intuitive. In this section and subsequent ones, we highlight some explicit use cases of *Sledgehammer* and *Nitpick*. They have been similarly applied at all levels as mentioned before.

3.2. Faithful Embedding of System E

It can be shown that the embedding of **E** in Isabelle/HOL is faithful [5], in the sense that a formula φ in the language of **E** is valid in the class PREF of all preference models if and only if the HOL translation of φ (notation: $\lfloor\varphi\rfloor$) is valid in the class of Henkin models of HOL.

Theorem 1 (Faithfulness of the embedding)**.**

$$\models^{\mathrm{PREF}} \varphi \text{ if and only if } \models^{\mathrm{HOL}} \lfloor\varphi\rfloor$$

Remember that the establishment of such a result is our main success criterium at layer L1 in the LogiKEy methodology.

This first two screenshots show the encoding of **E** in Isabelle/HOL. Fig. 2 shows the basic ingredients in the preferential model, and describes how the propositional and alethic modal connectives are handled. The betterness relation $\succeq$ is encoded as a binary relational constant r (l. 61). In Fig. 3, the notions of optimality and maximality are encoded. Different pairs of modal operators (obligation, permission) are introduced to distinguish between the two types of truth-conditions.

The model finder nitpick is able to verify the consistency of the formalization (l. 83) and to verify the non-equivalence between the two types of truth-conditions (l. 85). It is also able to

```
typedecl i (*Possible worlds.*)
type_synonym σ = "(i⇒bool)"
type_synonym α = "i⇒σ" (*Type of betterness relation between worlds.*)
type_synonym τ = "σ⇒σ"


consts aw::i (*Actual world.*)
abbreviation etrue  :: "σ" ("⊤") where "⊤ ≡ λw. True"
abbreviation efalse :: "σ" ("⊥")  where "⊥ ≡ λw. False"
abbreviation enot :: "σ⇒σ" ("¬_"[52]53)  where "¬φ ≡ λw. ¬φ(w)"
abbreviation eand :: "σ⇒σ⇒σ" (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w)∧ψ(w)"
abbreviation eor  :: "σ⇒σ⇒σ" (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w)∨ψ(w)"
abbreviation eimp :: "σ⇒σ⇒σ" (infixr"→"49) where "φ→ψ ≡ λw. φ(w)⟶ψ(w)"
abbreviation eequ :: "σ⇒σ⇒σ" (infixr"↔"48) where "φ↔ψ ≡ λw. φ(w)⟷ψ(w)"

abbreviation ebox :: "σ⇒σ" ("□") where "□ ≡ λφ w.  ∀v. φ(v)"
definition ddediomond  :: "σ⇒σ" ("◇") where "◇φ ≡ λw. ∃v. φ(v)"

abbreviation evalid :: "σ⇒bool" ("⌊_⌋"[8]109)  (*Global validity.*)
  where "⌊p⌋ ≡ ∀w. p w"
abbreviation ecjactual :: "σ⇒bool" ("⌊_⌋ı"[7]105) (*Local validity — in world aw.*)
  where "⌊p⌋ı ≡ p(aw)"

consts r :: "α" (infixr "r" 70) (*Betterness relation*)
```

Figure 2: Basic semantical ingredients; propositional and modal connectives

show the validity of the axioms of **E** and the invalidity of the axioms pertaining to the stronger systems under both evaluation rules.

4. Properties

The encoding of the properties of the betterness relation are shown in Fig. 4 and 5. On l. 106-116 of Fig. 4, one sees the different versions of Lewis' limit assumption. The property in Fig. 5 is the interval order condition. This one is usually described as the combination of totalness with the Ferrers condition as shown on l. 136. Sledgehammer is able to confirm a fact in general overlooked in the literature, that totalness can be replaced by the simpler condition of reflexivity (l. 139-141). More weakenings of transitivity are considered in the theory file. For simplicity's sake we put them aside.

5. Correspondences

5.1. Max rule

Here we check known correspondences between modal axioms under the max rule.

First, nitpick is able to confirm that the formula is not valid unless the matching property is assumed. Fig. 6 and 9 show that, when the relevant property is not assumed, a counter-model for D*, CM, DR and Sp is found by Nitpick.

```
65
66 abbreviation eopt  :: "σ⇒σ" ("opt<_>")  (* opt rule*)
67   where "opt<φ> ≡ (λv. ( (φ)(v) ∧ (∀x. ((φ)(x) ⟶ v r x) )) )"
68 abbreviation econdopt :: "σ⇒σ⇒σ" ("⊙<_|_>")
69   where "⊙<ψ|φ> ≡  λw. opt<φ> ⊆ ψ"
70 abbreviation eperm :: "σ⇒σ⇒σ" ("𝒫<_|_>")
71   where "𝒫<ψ|φ> ≡ ¬⊙<¬ψ|φ>"
72
73 abbreviation emax  :: "σ⇒σ" ("max<_>")      (*Max rule *)
74   where "max<φ> ≡ (λv. ( (φ)(v) ∧ (∀x. ((φ)(x) ⟶ (x r v ⟶  v r x)) )) )"
75 abbreviation econd  :: "σ⇒σ⇒σ" ("○<_|_>")
76   where "○<ψ|φ> ≡  λw. max<φ> ⊆ ψ"
77 abbreviation euncobl :: "σ⇒σ" ("O<_>")
78   where "O<φ> ≡ ○<φ|⊤>"
79 abbreviation ddeperm :: "σ⇒σ⇒σ" ("P<_|_>")
80   where "P<ψ|φ> ≡¬○<¬ψ|φ>"
81
82
83 lemma True nitpick [satisfy,user_axioms,expect=genuine] oops
84
85 lemma "⊙<ψ|φ> ≡ ○<ψ|φ>" nitpick [show_all] (*countermodel found*)
```

Figure 3: Truth-conditions

```
99 (*The standard properties*)
100 abbreviation reflexivity  where "reflexivity  ≡ (∀x. x r x)"
101 abbreviation transitivity
102   where "transitivity ≡ (∀x y z. (x r y ∧ y r z) ⟶ x r z)"
103 abbreviation totalness
104   where "totalness ≡ (∀x y. (x r y ∨ y r x))"
105
106 (*4 versions of Lewis's limit assumption*)
107 abbreviation mlimitedness
108   where "mlimitedness ≡ (∀φ. (∃x. (φ)x) ⟶ (∃x. max<φ>x))"
109 abbreviation msmoothness
110   where "msmoothness ≡ (∀φ x. ((φ)x ⟶
111                       (max<φ>x ∨ (∃y. (y r x ∧ ¬(x r y) ∧ max<φ>y)))))"
112 abbreviation olimitedness
113   where "olimitedness ≡ (∀φ. (∃x. (φ)x) ⟶ (∃x. opt<φ>x))"
114 abbreviation osmoothness  where
115    "osmoothness ≡ (∀φ x. ((φ)x ⟶
116                        (opt<φ>x ∨ (∃y. (y r x ∧ ¬(x r y) ∧ opt<φ>y)))))"
```

Figure 4: Standard properties

In Fig. 7, 8 and 9, it is confirmed that if the property is assumed, then the axiom is validated. Thus, the implications having the form "property ⇒ axiom" are all verified; Fig. 7 shows it for limitedness and smoothness, Fig. 8 for the interval order condition, and Fig. 9 for the combination of transitivity and totalness. But the converse implications are all falsified by Nitpick. We will come back to this point later on.

```
(*Interval order condition is totalness plus Ferrers*)
abbreviation Ferrers
  where "Ferrers ≡ (∀x y z u. ((x r u) ∧ (y r z)) ⟶ (x r z) ∨ (y r u))"

lemma assumes Ferrers reflexivity  (*fact overlooked in the literature*)
  shows totalness
  sledgehammer  (*proof found*)
```

Figure 5: Interval order

```
(*max-Limitedness corresponds to D*)
lemma "⌊◇φ → (○<ψ|φ> → P<ψ|φ>)⌋"
  nitpick [show_all]  (* counterexample found *)
  oops

lemma "⌊(○<ψ|φ>∧○<χ|φ>)→ ○<χ|φ∧ψ>⌋"
  nitpick [show_all] (* counterexample found *)
  oops

lemma "⌊○<χ|(φ∨ψ)>→((○<χ|φ>)∨(○<χ|ψ>))⌋"
  nitpick (* counterexample found *)
  oops
```

Figure 6: D*, CM and DR invalid in general

5.2. Opt rule

The outcomes of our experimentation are the same as for the max rule except for one small change. Transitivity no longer needs totalness to validate Sp. This one only needs transitivity. Besides the assumption of transitivity of the betterness relation gives us a principle of transitivity for a weak preference operator over formula, defined by $\phi \geq \psi$ iff $P(\phi/\phi \vee \psi)$. This is shown in Fig. 10.

5.3. Inclusion

In [1], proper inclusion between systems in the modal cube are verified by looking at the model constraints of their respective axiomatizations. Because of the lack of full equivalence between modal axiom and property of the relation, we cannot do the same, at least not yet. Nor can we show equivalence between systems when restraining the number of worlds.

6. The $\exists\forall$ truth-conditions (Lewis)

Variant evaluation rules have been proposed for the conditional in order to handle some of the problems encountered with the usual pattern of evaluation in terms of best. This section takes the example of Lewis [7]'s evaluation rule. In order to avoid commitment to the limit assumption, Lewis suggested that $\bigcirc(\psi/\varphi)$ should be true whenever there is no φ-world or there is a $\varphi \wedge \psi$-world which starts a (possibly infinite) sequence of increasingly better $\varphi \wedge \psi$-worlds. Formally:

```
lemma assumes "mlimitedness"
  shows  "D*": "⌊◇φ ⟶ ○<ψ|φ> ⟶ P<ψ|φ>⌋"
  sledgehammer
  oops (*proof found*)

lemma assumes "D*": "⌊◇φ ⟶ ¬(○<ψ|φ>∧○<¬ψ|φ>)⌋"
  shows "mlimitedness"
  sledgehammer (*all timed out*)
  nitpick [show_all] (* counterexample found *)
  oops

(*smoothness corresponds to cautious monotony *)
lemma assumes "msmoothness"
  shows  CM: "⌊(○<ψ|φ>∧○<χ|φ>)⟶ ○<χ|φ∧ψ>⌋"
  sledgehammer (*proof found*)
  oops

lemma assumes CM: "⌊(○<ψ|φ>∧○<χ|φ>)⟶ ○<χ|φ∧ψ>⌋"
  shows  "msmoothness"
  sledgehammer (* timed out*)
  nitpick [show_all] (* counterexample found *)
  oops
```

Figure 7: Limit assumption

$$a \vDash \bigcirc(\psi/\varphi) \text{ iff } \neg\exists b \ \ (b \models \varphi) \text{ or} \\ \exists b \ \ (b \models \varphi \wedge \psi \ \ \& \ \ \forall c \ (c \succeq b \Rightarrow c \models \varphi \rightarrow \psi)) \qquad (\exists\forall)$$

We shall refer to the statement appearing at the right-hand-side of "iff" as the $\exists\forall$ rule. The encoding is shown in Fig.11.

Isabelle/HOL is able to verify in what sense the standard account in terms of best requires the limit assumption. The law "from $\Diamond\varphi$, $\bigcirc(\psi/\varphi)$ and $\bigcirc(\neg\psi/\varphi)$ infer $\bigcirc(\chi/\varphi)$" is valid. This is known as the principle of "deontic explosion". It says that, in the presence of a conflict of duties (unless it is triggered by an "inconsistent" state of affairs) everything becomes obligatory. This has led most authors to make the limitedness assumption in order to validate D*, and hence make the principle of deontic explosion harmless: the set $\{\Diamond\varphi, \bigcirc(\psi/\varphi), \bigcirc(\neg\psi/\varphi)\}$ is not satisfiable. This is shown in Fig. 12. On l. 321, the validity of the DEX formula (=principle of deontic explosion) is shown under the max rule. On l. 326, the DEX formula is falsified under the $\exists\forall$ rule.

Isabelle/HOL is also able to verify that when all the standard properties of the betterness relation are assumed, then the three evaluation rules collapse. This is shown in Fig. 13. L.335 shows the equivalence between the $\exists\forall$ rule and the opt rule, and l. 342 shows the equivalence between the $\exists\forall$ rule and the max rule.

```
lemma assumes "reflexivity"
      (* assumes "Ferrers"*)
  shows  DR: "⌊○<χ|(φVψ)>⟶((○<χ|φ>)V(○<χ|ψ>))⌋"
  sledgehammer (* timed out*)
  nitpick (* counterexample found *)
  oops

lemma assumes "reflexivity" "Ferrers"
  shows  DR: "∀φ ψ χ.⌊○<χ|(φVψ)>⟶((○<χ|φ>)V(○<χ|ψ>))⌋"
  sledgehammer (* proof found *)
  nitpick (* no counterexample found *)
  oops

lemma assumes DR: "⌊○<χ|φVψ>⟶(○<χ|φ>V○<χ|ψ>)⌋"
  shows "reflexivity"
  sledgehammer (* timed out*)
  nitpick [show_all] (* counterexample found *)
  oops

lemma assumes DR: "⌊○<χ|φVψ>⟶(○<χ|φ>V○<χ|ψ>)⌋"
  shows "Ferrers"
  sledgehammer (* timed out*)
  nitpick (* counterexample found *)
  oops
```

Figure 8: Interval order

Questions of correspondence between properties and modal axioms are still under investigation. There are two extra complications. First, a completeness result is available for the strongest system **G** only: it is complete with respect to the class of models in which $\succeq$ is transitive and total (and hence reflexive). Second, only two properties seem to have an import, but the matching between them and the axioms is not one-to-one: one property validates more than one axiom, sometimes in combination with the other property. This is shown in Table 2. The left column gives the axiom. The right column shows the property (or pair of properties) required to validate this one.

Axiom of **G**	Property (or pair of properties) of $\succeq$
(D*)	totalness
(Sp)	transitivity
(COK)	transitivity and totalness
(CM)	transitivity and totalness

Table 2
Axioms and properties under the $\exists\forall$ rule (from [4])

In Fig. 14 Sledgehammer shows the validity of the axioms of **E** holding independently of the properties assumed of the betterness relation.

In Fig. 15 Sledgehammer confirms that the D* axiom and the Sp axiom call for totalness and transitivity, respectively.

Similarly, Fig. 16 shows that COK and CM call for *both* transitivity and totalness.

```
lemma assumes "transitivity"
  shows  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  sledgehammer (* timed out *)
  nitpick (* counterexample for card i=5*)
  oops

lemma assumes "totalness"
  shows  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  sledgehammer (* timed out*)
  nitpick (* counterexample for card i=4*)
  oops

lemma assumes "transitivity" "totalness"
  shows  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  sledgehammer (* proof found *)
  nitpick (* no counterexample found *)
  oops

lemma assumes  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  shows "totalness"
  sledgehammer (* timed out*)
  nitpick (* counterexample found for card i=1*)
  oops

lemma assumes  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  shows "transitivity"
  sledgehammer (* timed out*)
  nitpick (* counterexample found for card =4*)
```

Figure 9: Transitivity and totalness (max)

7. Discussion

To conclude, with regards to correspondence, the situation for conditional (deontic) logic is still slightly different from the one for traditional modal logic. In the latter setting, the full equivalence between the property of the relation and the modal formula is verified by automated means. In the former setting only the direction "property $\Rightarrow$ axiom" is verified by automated means. To be more precise, what is verified is the fact that, if the property holds, then the axiom holds. What is not confirmed is the converse statement, that if the axiom holds then the property holds. This asymmetry deserves to be discussed.

First, it is usual to distinguish between validity on a frame and validity in a model based on a frame. A frame is a pair $\mathcal{F} = (W, R)$, with W a set of worlds and R the accessibility relation. A model based on $\mathcal{F} = (W, R)$ is the triplet $\mathcal{M} = (W, R, V)$ obtained by adding a specific valuation V, or a specific assignment of truth-values to propositional letters at worlds. For a formula to be valid on a frame $\mathcal{F}$, it must be valid in all models based on $\mathcal{F}$. In other words, it must be true for every assignment to the propositional letters. We have worked at the level of models. But in so-called correspondence theory (see e.g. [19]) the link between formulas and properties is in general studied at the level of frames themselves. One shows that $\mathcal{F}$ meets a given condition iff formula A is valid on $\mathcal{F}$. In a recent extension of the semantical embedding approach for public announcement logic PAL, see [20], an explicit dependency on the concrete

```
lemma assumes "transitivity"
  shows  Sp: "⌊( 𝒫<ψ|φ> ∧ ⊙<(ψ⟶χ)|φ>) ⟶ ⊙<χ|(φ∧ψ)>⌋"
  sledgehammer (* proof found *)
  nitpick (* no counterexample found *)
  oops

lemma assumes "transitivity"
  shows  Trans: "⌊(𝒫<φ|φ∨ψ> ∧ 𝒫<ψ|ψ∨ξ>)⟶𝒫<φ|φ∨ξ>⌋"
  sledgehammer (* proof found *)
  nitpick [show_all] (* no counterexample found *)
  oops

lemma assumes Sp: "⌊( 𝒫<ψ|φ> ∧ ⊙<(ψ⟶χ)|φ>) ⟶ ⊙<χ|(φ∧ψ)>⌋"
      assumes Trans: "⌊(𝒫<φ|φ∨ψ> ∧ 𝒫<ψ|ψ∨ξ>)⟶𝒫<φ|φ∨ξ>⌋"
      shows "transitivity"
  sledgehammer (* timed out*)
  nitpick (* counterexample found for card i =3*)
  oops
```

Figure 10: Transitivity (opt)

```
abbreviation lewcond :: "σ⇒σ⇒σ"  ("○<_|_>")
  where "○<ψ|φ> ≡ λv. (¬(∃x. (φ)(x))∨
        (∃x. ((φ)(x)∧(ψ)(x) ∧ (∀y. ((y r x) ⟶ (φ)(y)⟶(ψ)(y))))))"
abbreviation lewperm :: "σ⇒σ⇒σ" ("∫<_|_>")
  where "∫<ψ|φ> ≡¬○<¬ψ|φ>"

lemma True nitpick [satisfy,user_axioms,expect=genuine]
_ oops
```

Figure 11: ∃∀ rule

```
(*deontic explosion-max rule*)
lemma DEX: "⌊(◇φ∧○<ψ|φ>∧○<¬ψ|φ>)⟶ ○<χ|φ>⌋"
  sledgehammer (*proof found*)
  oops

(*no-deontic explosion-lewis rule*)
lemma DEX: "⌊(◇φ∧○<ψ|φ>∧○<¬ψ|φ>)⟶ ○<χ|φ>⌋"
  sledgehammer (*timed out*)
  nitpick (*counter-model found for card i = 3*)
  oops
```

Figure 12: Deontic explosion (DEX)

evaluation domain has been modeled. It remains future work to study whether this idea can be further extended and adapted to also support a notion of validity for frames as needed here.

Second, the most we got is that a given property is a sufficient condition for the validity of the axiom, but not a necessary one. For instance, to disprove the implication "CM ⇒ m-smoothness" under the max rule (Fig. 7), Nitpick exhibits a model in which CM holds and m-smoothness falsified. The Henkin model is shown in Fig. 17. The corresponding preferential modal is also

```
lemma assumes "mlimitedness"
  assumes "transitivity"
  assumes "totalness"
  shows "⌊○<ψ|φ>↔⊙<ψ|φ>⌋"
  sledgehammer (*proof found*)
  oops

lemma assumes "mlimitedness"
  assumes "transitivity"
  assumes "totalness"
 shows "⌊○<ψ|φ>↔○<ψ|φ>⌋"
  sledgehammer (*proof found*)
  oops
```

Figure 13: Collapse

```
lemma Abs:"⌊○<ψ|φ> → □○<ψ|φ>⌋"
  sledgehammer (*proof found*)
  oops

lemma Nec:"⌊□ψ → ○<ψ|φ>⌋"
  sledgehammer (*proof found*)
  oops

lemma Ext:"⌊□(φ1↔φ2) → (○<ψ|φ1> ↔ ○<ψ|φ2>)⌋"
  sledgehammer (*proof found*)
  oops

lemma Id:"⌊○<φ|φ>⌋"
  sledgehammer (*proof found*)
  oops

lemma Sh:"⌊○<ψ|φ1∧φ2> → ○<(φ2→ψ)|φ1>⌋"
  sledgehammer (*proof found*)
  oops
```

Figure 14: Axioms independent of the properties (∃∀ rule)

shown in the box. An arrow from i_1 to i_2 means $i_1 \succeq i_2$. No arrow from i_2 to i_1 means $i_2 \not\succeq i_1$. Smoothness is falsified, because it contains an infinite loop of strict betterness, making the smoothness condition fail for, e.g., $\varphi \vee \neg\varphi$. But CM (vacuously) holds, because the two conjuncts appearing in the antecedent of the axiom are both false. Indeed, i_3 is a maximal φ-world, and it falsifies ψ and χ. This shows that m-smoothness is not a necessary condition for the axiom to hold.

It is interesting to remark that, because a counter-model generated by Nitpick is always finite, this Henkin model is also a standard one. We leave it as a topic for future research to investigate if the crucial distinction between standard and non-standard models, which (according to Andrews [21]) sheds so much light on the mysteries associated with the incompleteness theorems, has a bearing on the issue at hand.

Another open problem concerns the possibility of verifying "negative" results. As shown in Table 1, under the max rule transitivity alone does not correspond to any axiom. Also under

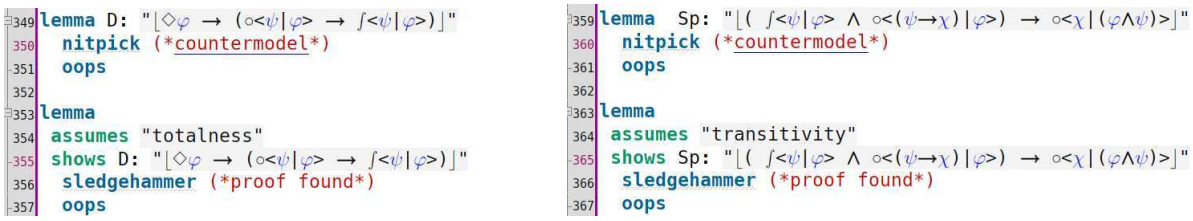

Figure 15: Transitivity and totalness alone (∃∀ rule)

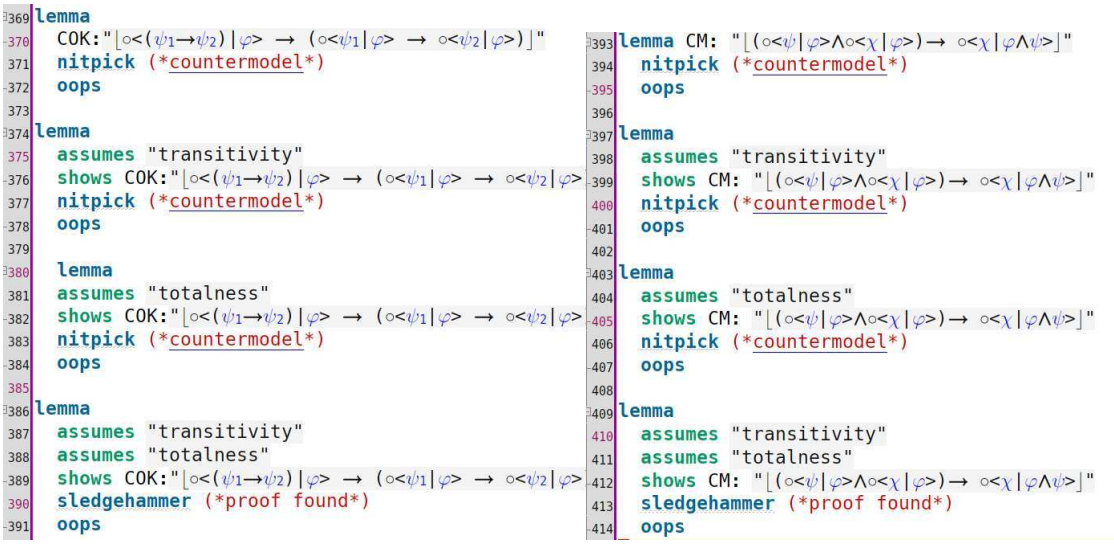

Figure 16: Transitivity and totalness together (∃∀ rule)

both the max rule and the opt rule neither reflexivity nor totalness correspond to an axiom. Finally, under the ∃∀ rule the limit assumption has no import.All this has been established with pen and paper. It would be worth exploring the question as to whether and how this problem could be tackled in Isabelle/HOL.

Acknowledgments

Xavier Parent was funded in whole, or in part, by the Austrian Science Fund (FWF) [M 3240-N, ANCoR project]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We thank the anonymous reviewers for their valuable comments which helped to improve this paper.

```
Nitpicking formula...
Kodkod warning: cannot launch SAT solver, falling back on "DefaultSAT4J
Nitpick found a counterexample for card i = 3:

  Free variables:
    χ = (λx. _)(i₁ := False, i₂ := True, i₃ := False)
    φ = (λx. _)(i₁ := False, i₂ := False, i₃ := True)
    ψ = (λx. _)(i₁ := False, i₂ := False, i₃ := False)
  Skolem constants:
    φ = (λx. _)(i₁ := True, i₂ := True, i₃ := True)
    x = i₃
    x = i₂
    λy. x = (λx. _)(i₁ := i₃, i₂ := i₁, i₃ := i₃)
  Constant:
    (r) =
      (λx. _)
      (i₁ := (λx. _)(i₁ := False, i₂ := True, i₃ := False),
       i₂ := (λx. _)(i₁ := False, i₂ := True, i₃ := True),
       i₃ := (λx. _)(i₁ := True, i₂ := False, i₃ := False))
```

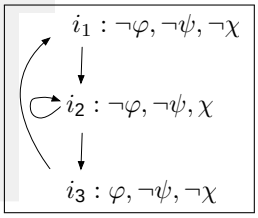

Figure 17: A non-smooth model validating CM (max).

References

[1] C. Benzmüller, M. Claus, N. Sultana, Systematic verification of the modal logic cube in Isabelle/HOL, in: C. Kaliszyk, A. Paskevich (Eds.), PxTP 2015, volume 186, EPTCS, Berlin, Germany, 2015, pp. 27–41.

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

[3] X. Parent, L. van der Torre, Introduction to Deontic Logic and Normative Systems, College Publications, London, 2021.

[4] X. Parent, Preference semantics for dyadic deontic logic: a survey of results, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden, L. van der Torre (Eds.), Handbook of Deontic Logic and Normative Systems, volume 2, College Publications, London. UK, 2021, pp. 7–70.

[5] C. Benzmüller, A. Farjami, X. Parent, Åqvist's dyadic deontic logic E in HOL, Journal of Applied Logics – IfCoLoG Journal of Logics and their Applications (Special Issue: Reasoning for Legal AI) 6 (2019) 733–755.

[6] G. E. Hughes, M. J. Cresswell, A companion to modal logic, Methuen, London, 1984.

[7] D. Lewis, Counterfactuals, Blackwell, Oxford, 1973.

[8] R. Luce, Semiorders and a theory of utility discrimination, Econometrica 24 (1956) 178–191.

[9] X. Parent, Maximality vs. optimality in dyadic deontic logic, Journal of Philosophical Logic 43 (2014) 1101–1128.

[10] B. Chellas, Basic conditional logic, Journal of Philosophical Logic 4 (1975) pp. 133–153.

[11] X. Parent, Completeness of Åqvist's systems E and F, Review of Symbolic Logic 8 (2015) 164–177.

[12] C. Benzmüller, D. Gabbay, V. Genovese, D. Rispoli, Embedding and automating conditional logics in classical higher-order logic, Annals of Mathematics and Artificial Intelligence 66 (2012) 257–271.

[13] X. Parent, On some weakened forms of transitivity in the logic of norms, in: L. Giordanni, G. Casini (Eds.), NMR 2022: 20th International Workshop on Non-Monotonic Reasoning, CEUR-WS, 2022, pp. 147–150. Extended abstract.

[14] C. Benzmüller, X. Parent, L. van der Torre, Designing normative theories for ethical and

legal reasoning: Logikey framework, methodology, and tool support, Artificial Intelligence 287 (2020) 103348.
[15] C. Benzmüller, Universal (meta-)logical reasoning: Recent successes, Science of Computer Programming 172 (2019) 48–62.
[16] J. C. Blanchette, S. Böhme, L. C. Paulson, Extending Sledgehammer with SMT solvers, Journal of Automated Reasoning 51 (2013) 109–128.
[17] J. C. Blanchette, C. Kaliszyk, L. C. Paulson, J. Urban, Hammering towards QED, Journal of Formalized Reasoning 9 (2016) 101–148.
[18] J. C. Blanchette, T. Nipkow, Nitpick: A counterexample generator for higher-order logic based on a relational model finder, in: M. Kaufmann, L. C. Paulson (Eds.), ITP 2010, volume 6172 of *LNCS*, Springer, 2010, pp. 131–146.
[19] J. Van Benthem, Correspondence theory, in: D. M. Gabbay, F. Guenthner (Eds.), Handbook of Philosophical Logic, Springer Netherlands, Dordrecht, 2001, pp. 325–408.
[20] C. Benzmüller, S. Reiche, Automating public announcement logic with relativized common knowledge as a fragment of HOL in LogiKEy, Journal of Logic and Computation (2022).
[21] P. Andrew, An introduction to mathematical logic and type theory, Springer, 2002.