Normative Conditional Reasoning as a Fragment of HOL

Xavier Parent[a] and Christoph Benzmüller[b] [c]

[a] Technische Universität Wien, Favoritenstrasse 9, A-1040 Wien, Austria [b] Universität Bamberg, An der Weberei 5, 96047 Bamberg, Germany [c] Freie Universität Berlin, Arnimallee 7, 14195 Berlin, Germany

ARTICLE HISTORY
Compiled August 23, 2023

ABSTRACT

We report some results regarding the mechanization of normative (preference-based) conditional reasoning. Our focus is on Åqvist's system **E** for conditional obligation (and its extensions). Our mechanization is achieved via a shallow semantical embedding in Isabelle/HOL. We consider two possible uses of the framework. The first one is as a tool for meta-reasoning about the considered logic. We employ it for the automated verification of deontic correspondences (broadly conceived) and related matters, analogous to what has been previously achieved for the modal logic cube. The second use is as a tool for assessing ethical arguments. We provide a computer encoding of a well-known paradox in population ethics, Parfit's repugnant conclusion. Whether the presented encoding increases or decreases the attractiveness and persuasiveness of the repugnant conclusion is a question we would like to pass on to philosophy and ethics.

KEYWORDS
Conditional obligation; betterness; Isabelle/HOL; automated theorem proving; population ethics; mere addition/repugnant conclusion paradox

1. Introduction

We report some results regarding the mechanization of normative (preference-based) conditional reasoning. Our focus is on Åqvist's system **E** for conditional obligation (and its extensions). Our mechanization is achieved via a shallow semantical embedding in Isabelle/HOL adapting the methods used by Benzmüller et al. (2015). To look at Standard Deontic Logic (SDL) and extensions (Chellas, 1980; Parent and van der Torre, 2021) would not be very interesting. First, no new insights would be gained, since SDL is a normal modal logic of type KD, which is already covered by the prior work of Benzmüller et al. Second, SDL is vulnerable to the well-known deontic paradoxes, including in particular Chisholm's paradox of contrary-to-duty obligation, see Parent and van der Torre (2021) for details. We thus focus here on Dyadic Deontic Logics (DDL) with a preference-based semantics, which originate from the works of Hansson (1969) and Lewis (1973). In these works, one uses an "intensional" conditional to represent conditional obligation sentences that is weaker than the one obtained us-

CONTACT X. Parent. Email: x.parent.xavier@gmail.com; C. Benzmüller: christoph.benzmueller@uni-bamberg.de

ing material implication. The semantics generalizes that of SDL: the SDL-ish binary classification of states into good/bad is relaxed to allow for grades of ideality (best, second-best, ...). In this framework, which is particularly popular in deontic logic (cf. the overview chapter by Parent (2021) in the second volume of the *Handbook of Deontic Logic*), 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.

In this paper, our emphasis is on two possible uses of the mechanized tool. First, we employ it as a tool for meta-reasoning about the considered logics. So far the correspondences between properties and modal axioms have been established "with pen and paper". This raises the question of how much of these correspondences can be automatically explored by modern theorem-proving technology. The automatic verification of correspondences can be done for the modal cube (Benzmüller et al., 2015). We want to understand if it can also be done in DDL. As explained by Benzmüller et al. (2015) we believe that "automation facilities could be very useful for the exploration of 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 further prior results from Benzmüller et al. (2019), 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**. 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." (Hughes and Cresswell, 1984, p. 41)

The second use we consider for our mechanized system is as a tool for assessing ethical arguments in philosophical debates. As an illustration, we look at one of the well-known impossibility theorems in population ethics, the so-called "repugnant conclusion" due to Parfit (1984). We provide a computer encoding of the repugnant conclusion in order to make it amenable to formal analysis and computer-assisted experiments. We believe that the formalisation has the potential to further stimulate the philosophical debate on the repugnant conclusion, since the simplifications achieved are indeed quite far-reaching. In particular, our formalization indicates the possibility of a new take on the scenario. It has been suggested that, since the transitivity of "better than" is presupposed in the impossibility theorem, one can avoid such a result by simply giving up on transitivity. The solution is sometimes dismissed on the ground that it is too radical. The formalization reveals a less radical variant solution, which consists in weakening transitivity rather than giving it up wholesale. However, not all candidate weakenings of transitivity will do. For instance, a-cyclicity works fine, but not the interval order condition.

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 studies the corre-

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

spondence between the properties of the betterness relation and the axioms. Section 5 discusses the repugnant conclusion. Section 6 concludes.[2]

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 $\forall b \in \text{best}(\varphi)$ 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**
$\text{best}(\varphi) = \max(\varphi)$	$\text{best}(\varphi) = \text{opt}(\varphi)$

where

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

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);

[2]The theory file is available for downloading at http://logikey.org under sub-repository "/Deontic-Logics/cube-dll/" (files "cube.thy" and "mere_addition.thy").

- Interval order: $\succeq$ is reflexive and Ferrers (if $a \succeq b$ and $c \succeq d$, then $a \succeq d$ or $c \succeq b$);
- A-cyclicity: $\succ$ contains no cycles of the form $a_1 \succ a_2 \succ ... \succ a_n \succ a_1$.

A-cyclicity and the interval order condition are discussed in Parent (2022). They are two weakenings of the assumption of transitivity. In particular, 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 Luce 1956).

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 best $\in \{\max, \text{opt}\}$

Limitedness

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

Smoothness (or stopperedness)

If $x \models \varphi$, then: either $x \in \text{best}(\varphi)$ or $\exists y$ s.t. $y \succ x$ & $y \in \text{best}(\varphi)$ (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 the literature, and the relationships between these versions of the limit assumption, see Parent (2014).

The above semantics may be viewed as a special case of the selection function semantics favored by Stalnaker and generalized by Chellas (1975). 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 thus given as: $\bigcirc(\psi/\varphi)$ 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− cf. Parent (2015) for details.[3] The correspondence between constraints put on the selection function and modal axioms have been verified by automated means by Benzmüller et al. (2012). A comparison between this prior study and ours 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.

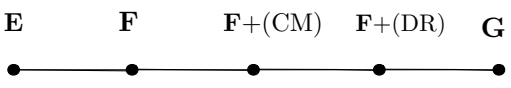


Figure 1.: Systems

[3]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 (cf. Chellas (1975)).

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

- For **E** (the naming follows Parent (2021)):

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

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

$$\Diamond\varphi \to (\bigcirc(\psi/\varphi) \to P(\psi/\varphi)) \tag{D*}$$

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

$$(\bigcirc(\psi/\varphi) \wedge \bigcirc(\xi/\varphi)) \to \bigcirc(\xi/\varphi \wedge \psi) \tag{CM}$$

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

$$\bigcirc(\xi/\varphi \vee \psi) \to (\bigcirc(\xi/\varphi) \vee \bigcirc(\xi/\psi)) \tag{DR}$$

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

$$(P(\psi/\varphi) \wedge \bigcirc(\psi \to \xi/\varphi)) \to \bigcirc(\xi/\varphi \wedge \psi) \tag{Sp}$$

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 Lewis (1973), and reflects the fact that the ranking is not world-relative. Nec is the deontic counterpart of the familiar necessitation rule. 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.

We give below the main soundness and completeness theorems. Those stated in Th. 2.1 hold under both the opt rule and the max rule. It is understood that limitedness is cast in terms of opt when the opt rule is applied, and in terms of max when the max rule is applied. The same holds for smoothness.

Theorem 2.1 (Soundness and completeness, Parent2021; 2022)**.** *(i)* **E** *is sound and complete w.r.t. the class of all preference models; (ii)* **F** *is sound and complete w.r.t. the class of preference models in which $\succeq$ is limited; (iii)* **F**$+CM$ *is sound and complete w.r.t. the class of preference models in which $\succeq$ is smooth; (iv)* **F**$+DR$ *is (weakly) sound and complete w.r.t. the class of (finite) preference models in which $\succeq$ meets the interval order condition.*

Theorem 2.2 (Soundness and completeness, Parent2014)**.** *(i) Under the opt-rule* **G** *is sound and complete w.r.t. the class of preference models in which* $\succeq$ *is limited and transitive; (ii) under the max-rule,* **G** *is sound and complete w.r.t. the class of preference models in which* $\succeq$ *is limited, transitive and total.*

For more background on these systems, see Parent (2021) and the references therein.

2.3. Correspondences

Table 1 shows some of the known "correspondences" between semantic properties and formulas that can be extracted from Th. 2.1 and Th. 2.2. The leftmost column shows the properties of $\succeq$. The two middle columns show the corresponding modal axioms, the first column for the max rule, and the second one for the opt rule. As before 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, in the sense that the property does not modify the set of valid formulas. On the fifth (resp. seventh) line the parenthesis "(+smoothness)" (resp. "(+limitedness)") indicates that smoothness (resp. limitedness) is assumed in the background.[4]

Property	Formula (max)	Formula (opt)	Reference
reflexivity	$\times$	$\times$	Parent (2015)
totalness	$\times$	$\times$	Parent (2015)
limitedness	$D^\star$	$D^\star$	Parent (2015)
smoothness	CM	CM	Parent (2014)
transitivity (+smoothness)	$\times$	Sp	Parent (2014,2)
transitivity+totalness	Sp	$\times$	Parent (2014)
interval order (+limitedness)	DR	DR	Parent (2022)

Table 1.: Some correspondences

3. System E in Isabelle/HOL

Our modelling of System **E** in Isabelle/HOL reuses and adapts prior work (Benzmüller et al., 2019) and it instantiates and applies the LogiKEy methodology (Benzmüller et al., 2020), which supports plurality at different modelling layers.

3.1. LogiKEy

Classical higher-order logic (HOL) is fixed in the LogiKEy methodology and infrastructure (Benzmüller et al., 2020) as a *universal meta-logic* (Benzmüller, 2019) at the base layer (L0), on top of which a plurality of (combinations of) object logics can

[4]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.

```
theory DDLcube imports Main

begin

(*** We introduce Aqvist's system E from the 2019 IfColog paper ***)

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 eimpf :: "σ⇒σ⇒σ" (infixr"→"49) where "φ→ψ ≡ λw. φ(w)⟶ψ(w)"
abbreviation eimpb :: "σ⇒σ⇒σ" (infixr"←"49) where "φ←ψ ≡ λw. ψ(w)⟶φ(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)"
abbreviation 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" ("⌊_⌋ι"[7]105) (* Local validity — in world aw *)
  where "⌊p⌋ι ≡ p(aw)"

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

Figure 2.: Basic semantical ingredients; propositional and modal connectives

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 as *Sledgehammer* (Blanchette et al., 2013; Blanchette et al., 2016) and *Nitpick* (Blanchette and Nipkow, 2010). 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.

```
36 abbreviation eopt  :: "σ⇒σ" ("opt<_>")  (* opt rule*)
37   where "opt<φ> ≡ (λv. ( (φ)(v) ∧ (∀x. ((φ)(x) ⟶ v r x) )) )"
38 abbreviation econdopt :: "σ⇒σ⇒σ" ("⊙<_|_>")
39   where "⊙<ψ|φ> ≡  λw. opt<φ> ⊆ ψ"
40 abbreviation eperm :: "σ⇒σ⇒σ" ("𝒫<_|_>")
41   where "𝒫<ψ|φ> ≡ ¬⊙<¬ψ|φ>"
42
43 abbreviation emax  :: "σ⇒σ" ("max<_>")      (* Max rule *)
44   where "max<φ> ≡ (λv. ( (φ)(v) ∧ (∀x. ((φ)(x) ⟶ (x r v ⟶  v r x)) )) )"
45 abbreviation econd  :: "σ⇒σ⇒σ" ("○<_|_>")
46   where "○<ψ|φ> ≡  λw. max<φ> ⊆ ψ"
47 abbreviation euncobl :: "σ⇒σ" ("O<_>")
48   where "O<φ> ≡ ○<φ|⊤>"
49 abbreviation ddeperm :: "σ⇒σ⇒σ" ("P<_|_>")
50   where "P<ψ|φ> ≡¬○<¬ψ|φ>"
51
52
53 (* Settings for model finder Nitpick *)
54
55 nitpick_params [user_axioms,show_all,expect=genuine]
56
57
58 (*** First consistency check ***)
59
60 lemma True
61   nitpick [satisfy] (* model found *)
62   oops
63
64 (*** The max-rule and opt-rule don't coincide ***)
65
66 lemma "⊙<ψ|φ> ≡ ○<ψ|φ>"
67   nitpick [card i=1] (* counterexample found for card i=1 *)
68   oops
```

Figure 3.: Truth-conditions

3.2. Faithful embedding of system **E**

In the work of Benzmüller et al. (2019), it is shown that the embedding of **E** in Isabelle/HOL is faithful, 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 3.1 (Faithfulness of the embedding)**.**

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

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. 31). 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. 60) and to verify the non-equivalence between the two types of truth-conditions (l. 66). *Sledgehammer* is able to show the validity of the axioms of **E**. By presenting a suitable counter-model, *Nitpick* is able to show the invalidity of the axioms pertaining to the stronger systems.

3.3. Properties

The encoding of the properties of the betterness relation are shown in Figs. 4 and 5. On l. 99-104 of Fig. 4, one sees the different versions of Lewis' limit assumption. The

```
(****************
Properties
******************)

(* The standard properties of the betterness relation *)

abbreviation "reflexivity  ≡ (∀x. x r x)"
abbreviation "transitivity ≡ (∀x y z. (x r y ∧ y r z) ⟶ x r z)"
abbreviation "totalness ≡ (∀x y. (x r y ∨ y r x))"

(* 4 versions of Lewis's limit assumption *)

abbreviation "mlimitedness ≡ (∀φ. (∃x. (φ)x) ⟶ (∃x. max<φ>x))"
abbreviation "msmoothness ≡
   (∀φ x. ((φ)x ⟶ (max<φ>x ∨ (∃y. (y r x ∧ ¬(x r y) ∧ max<φ>y)))))"
abbreviation "olimitedness ≡ (∀φ. (∃x. (φ)x) ⟶ (∃x. opt<φ>x))"
abbreviation "osmoothness ≡
   (∀φ x. ((φ)x ⟶ (opt<φ>x ∨ (∃y. (y r x ∧ ¬(x r y) ∧ opt<φ>y)))))"
```

Figure 4.: Standard properties

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. 137. *Sledgehammer* is able to confirm a fact that has been generally overlooked in the literature, namely that totalness can be replaced by the simpler condition of reflexivity (l. 139-141). More

```
(* Interval order (reflexivity + Ferrers) *)

abbreviation Ferrers
  where "Ferrers ≡ (∀x y z u. (x r u ∧ y r z) ⟶ (x r z ∨ y r u))"

theorem T2:
  assumes Ferrers and reflexivity  (*fact overlooked in the literature*)
  shows totalness
  by (simp add: assms(1) assms(2))  (* proof found *)
```

Figure 5.: Interval order

weakenings of transitivity Parent (2022) are encoded in the theory file. For simplicity's sake we give the example of quasi-transitivity and a-cyclicity. The encoding of the second is shown in Fig. 6. The formal definition of this notion reads: if $a \succ^\star b$, where $\succ^\star$ is the transitive closure of $\succ$, then $b \not\succ a$. In Isabelle/HOL, the transitive closure of a relation can be defined in a few lines, shown in Fig. 7.

```
(* A-cyclicity: cycles of strict betterness are ruled out*)

abbreviation loopfree
  where "loopfree ≡ ∀x y. tcr_strict x y ⟶ (y r x ⟶ x r y)"
```

Figure 6.: A-cyclicity

Quasi-transitivity requires the strict betterness relation be transitive. Its encoding is shown in Fig. 8.

```
(* Weaker forms of transitivity--they require the notion of
transitive closure*)

definition transitive :: "α⇒bool"
  where "transitive Rel ≡ ∀x y z. Rel x y ∧  Rel y z ⟶ Rel x z"
definition sub_rel :: "α⇒α⇒bool"
  where "sub_rel Rel1 Rel2 ≡ ∀u v. Rel1 u v ⟶ Rel2 u v"
definition assfactor::"α⇒α"
  where "assfactor Rel ≡ λu v. Rel u v ∧ ¬Rel v u "

(* In HOL the transitive closure of a relation can be defined in a single line; here
   we apply the construction to betterness relation r and for its strict
   variant (λu v. u r v ∧ ¬v r u) *)
definition tcr
  where "tcr ≡ λx y. ∀Q. transitive Q ⟶ (sub_rel r Q ⟶ Q x y)"

definition tcr_strict
  where "tcr_strict ≡ λx y. ∀Q. transitive Q ⟶ (sub_rel (λu v. u r v ∧ ¬v r u) Q ⟶ Q x y)"
```

Figure 7.: Transitive closure

```
(* Quasi-transitivity: the strict betterness
relation is transitive*)
abbreviation Quasitransit
  where "Quasitransit  ≡ ∀x y z. (assfactor r x y ∧
                    assfactor r y z) ⟶ assfactor r x z"
```

Figure 8.: Quasi-transitivity

4. Correspondences

4.1. 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. Figs. 9 and 12 show that, when the relevant property is not assumed, counter-models for D*, CM, DR and Sp are found by *Nitpick*.

```
(* Max-Limitedness corresponds to D *)

lemma "⌊◇φ ⟶ (○<ψ|φ> ⟶ P<ψ|φ>)⌋"
  nitpick [card i=3]  (* counterexample found for card i=3 *)
  oops

lemma "⌊(○<ψ|φ> ∧ ○<χ|φ>) ⟶ ○<χ|φ∧ψ>⌋"
  nitpick [card i=3]  (* counterexample found *)
  oops

lemma "⌊○<χ|(φ∨ψ)> ⟶ ((○<χ|φ>) ∨ (○<χ|ψ>))⌋"
  nitpick [card i=3]  (* counterexample found *)
  oops
```

Figure 9.: D*, CM and DR invalid in general

In Figs. 10, 11 and 12, 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. 10 shows it for limitedness and smoothness, Fig. 11 for the interval order condition, and Fig. 12 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.

```
theorem T8:
  assumes mlimitedness
  shows  "D*": "⌊◇φ ⟶ ○<ψ|φ> ⟶ P<ψ|φ>⌋"
  sledgehammer
  using assms
  oops

lemma
  assumes "D*": "⌊◇φ ⟶ ¬(○<ψ|φ> ∧ ○<¬ψ|φ>)⌋"
  shows mlimitedness
  nitpick [card i=3]  (* counterexample found *)
  oops

(* Smoothness corresponds to cautious monotony *)

theorem T9:
  assumes msmoothness
  shows CM: "⌊(○<ψ|φ> ∧ ○<χ|φ>) ⟶ ○<χ|φ∧ψ>⌋"
  using assms by force

lemma
  assumes CM: "⌊(○<ψ|φ> ∧ ○<χ|φ>) ⟶ ○<χ|φ∧ψ>⌋"
  shows  msmoothness
  nitpick [card i=3]  (* counterexample found *)
  oops
```

Figure 10.: Limit assumption

```
lemma
  assumes reflexivity
  shows  DR: "⌊○<χ|φ∨ψ> ⟶ (○<χ|φ> ∨ ○<χ|ψ>)⌋"
  nitpick [card i=3]  (* counterexample found *)
  oops

theorem T10:
  assumes reflexivity and Ferrers
  shows  DR: "⌊○<χ|(φ∨ψ)> ⟶ (○<χ|φ> ∨ ○<χ|ψ>)⌋"
  by (metis assms)

lemma
  assumes DR: "⌊○<χ|φ∨ψ> ⟶ (○<χ|φ> ∨ ○<χ|ψ>)⌋"
  shows reflexivity
  nitpick [card i=1]  (* counterexample found *)
  oops

lemma
  assumes DR: "⌊○<χ|φ∨ψ>⟶(○<χ|φ> ∨ ○<χ|ψ>)⌋"
  shows Ferrers
  nitpick [card i=2]  (* counterexample found *)
  oops
```

Figure 11.: Interval order

4.2. 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 $\varphi \geq \psi$ iff $P(\varphi/\varphi \vee \psi)$. This is shown in Fig. 13.

```
(*Transitivity and totalness corresponds to the Spohn axiom (Sp)*)

lemma
  assumes transitivity
  shows  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ→χ)|φ>) → ○<χ|(φ∧ψ)>⌋"
  nitpick [card i=3] (* counterexample found *)
  oops

lemma
  assumes totalness
  shows  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ→χ)|φ>) → ○<χ|(φ∧ψ)>⌋"
  nitpick [card i=3] (* counterexample *)
  oops

theorem T11:
  assumes transitivity and totalness
  shows  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ→χ)|φ>) → ○<χ|(φ∧ψ)>⌋"
  by (metis assms)

theorem T12:
  assumes transitivity and totalness
  shows  transit: "⌊( P<φ|φ∨ψ> ∧ P<ψ|ψ∨χ>) → P<φ|(φ∨χ)>⌋"
  by (metis assms(1) assms(2))

lemma
  assumes  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ→χ)|φ>) → ○<χ|(φ∧ψ)>⌋"
  shows totalness
  nitpick [card i=1] (* counterexample found *)
  oops

lemma
  assumes  Sp: "⌊( P<ψ|φ> ∧ ○<(ψ→χ)|φ>) → ○<χ|(φ∧ψ)>⌋"
  shows transitivity
  nitpick [card i=2] (* counterexample found *)
  oops
```

Figure 12.: Transitivity and totalness (max)

```
(*transitivity*)

theorem T15:
  assumes transitivity
  shows  Sp': "⌊( 𝒫<ψ|φ> ∧ ⊙<(ψ→χ)|φ>) → ⊙<χ|(φ∧ψ)>⌋"
  by (metis assms)

theorem T16:
  assumes transitivity
  shows  Trans: "⌊( 𝒫<φ|φ∨ψ> ∧ 𝒫<ψ|ψ∨ξ> )→ 𝒫<φ|φ∨ξ>⌋"
  by (metis assms)

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
  nitpick [card i=2] (* counterexample found  *)
  oops
```

Figure 13.: Transitivity (opt)

4.3. Inclusion

In the work of Benzmüller et al. (2015), 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.

4.4. The ∃∀ 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. We take the example of Lewis (1973)'s evaluation rule. In order to avoid commitment to the limit assumption, Lewis suggests 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:

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

We shall refer to the statement appearing at the right-hand-side of "iff" as the ∃∀ rule. The encoding is shown in Fig.14.

```
89 abbreviation lewcond :: "σ⇒σ⇒σ"  ("○<_|_>")
90   where "○<ψ|φ> ≡ λv. (¬(∃x. (φ)(x))∨
91         (∃x. ((φ)(x)∧(ψ)(x) ∧ (∀y. ((y r x) ⟶ (φ)(y)⟶(ψ)(y))))))"
92 abbreviation lewperm :: "σ⇒σ⇒σ" ("∫<_|_>")
93   where "∫<ψ|φ> ≡¬○<¬ψ|φ>"
94
95 lemma True nitpick [satisfy,user_axioms,expect=genuine]
96   oops
```

Figure 14.: ∃∀ rule

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", often called DEX. 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 DEX harmless: the set $\{\Diamond\varphi, \bigcirc(\psi/\varphi), \bigcirc(\neg\psi/\varphi)\}$ is not satisfiable. This is shown in Fig. 15. On l. 371, the validity of DEX is established under the max rule. On l. 374, DEX is falsified under the ∃∀ 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. 15. L. 380-384 show the equivalence between the ∃∀ rule and the opt rule, and l. 386-390 show the equivalence between the ∃∀ rule and the max rule.

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.

```
(****************
Relationship Lewis rule and max/opt rule
*****************)

(* deontic explosion-max rule *)
theorem DEX: "⌊(◇φ ∧ ○<ψ|φ> ∧ ○<¬ψ|φ>) → ○<χ|φ>⌋"
  by blast

(* no-deontic explosion-lewis rule *)
lemma DEX: "⌊(◇φ ∧ ○<ψ|φ> ∧ ○<¬ψ|φ>) → ○<χ|φ>⌋"
  nitpick [card i=2] (* counterexample  found*)
  oops

theorem T18:
  assumes mlimitedness and transitivity and totalness
  shows "⌊○<ψ|φ>↔⊙<ψ|φ>⌋"
  sledgehammer
  by (smt (z3) assms)

theorem T19:
  assumes mlimitedness and transitivity and totalness
  shows "⌊○<ψ|φ>↔○<ψ|φ>⌋"
  sledgehammer
  by (smt (z3) assms)
```

Figure 15.: Deontic explosion (DEX)

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 ∃∀ rule—from Parent (2021)

```
(*axioms of E holding irrespective of the properties of r*)

theorem Abs: "⌊○<ψ|φ> → □○<ψ|φ>⌋"
  by blast

theorem Nec: "⌊□ψ → ○<ψ|φ>⌋"
  by blast

theorem Ext: "⌊□(φ1↔φ2) → (○<ψ|φ1> ↔ ○<ψ|φ2>)⌋"
  by simp

theorem Id: "⌊○<φ|φ>⌋"
  by auto

theorem  Sh: "⌊○<ψ|φ1∧φ2> → ○<(φ2→ψ)|φ1>⌋"
  by blast
```

Figure 16.: Axioms independent of the properties (∃∀ rule)

```
(* axioms of E holding if r transitive and total *)
lemma D: "⌊◇φ ⟶ (○<ψ|φ> ⟶ ∫<ψ|φ>)⌋"
  nitpick [card i=2] (* counterexample found *)
  oops

theorem T20:
  assumes totalness
  shows D': "⌊◇φ ⟶ (○<ψ|φ> ⟶ ∫<ψ|φ>)⌋"
  by (metis assms)
```

Figure 17.: Transitivity and totalness alone (∃∀ rule)

```
lemma  Sp: "⌊( ∫<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  nitpick (*countermodel*)
  oops

lemma
 assumes "transitivity"
 shows Sp: "⌊( ∫<ψ|φ> ∧ ○<(ψ⟶χ)|φ>) ⟶ ○<χ|(φ∧ψ)>⌋"
  sledgehammer (*proof found*)
  oops
```

Figure 18.: Transitivity and totalness alone (∃∀ rule, ct'd)

```
lemma COK:"⌊○<(ψ1⟶ψ2)|φ> ⟶ (○<ψ1|φ> ⟶ ○<ψ2|φ>)⌋"
  nitpick [card i=2] (* counterexample found *)
  oops

lemma
  assumes transitivity
  shows COK:"⌊○<(ψ1⟶ψ2)|φ> ⟶ (○<ψ1|φ> ⟶ ○<ψ2|φ>)⌋"
  nitpick [card i=2] (* counterexample found *)
  oops

lemma
  assumes totalness
  shows COK:"⌊○<(ψ1⟶ψ2)|φ> ⟶ (○<ψ1|φ> ⟶ ○<ψ2|φ>)⌋"
  nitpick [card i=3] (* counterexample found *)
  oops

theorem T22:
  assumes transitivity and totalness
  shows COK:"⌊○<(ψ1⟶ψ2)|φ> ⟶ (○<ψ1|φ> ⟶ ○<ψ2|φ>)⌋"
 by (smt (z3) assms)
```

Figure 19.: Transitivity and totalness together (∃∀ rule)

```
lemma CM: "⌊(○<ψ|φ>∧○<χ|φ>)⟶ ○<χ|φ∧ψ>⌋"
  nitpick [card i=2] (* counterexample found *)
  oops

lemma
  assumes transitivity
  shows CM: "⌊(○<ψ|φ>∧○<χ|φ>)⟶ ○<χ|φ∧ψ>⌋"
  nitpick [card i=2] (* counterexample found *)
  oops

lemma
  assumes totalness
  shows CM: "⌊(○<ψ|φ>∧○<χ|φ>)⟶ ○<χ|φ∧ψ>⌋"
  nitpick [card i=3] (* counterexample found *)
  oops

theorem T23:
  assumes transitivity and totalness
  shows CM'': "⌊(○<ψ|φ>∧○<χ|φ>)⟶ ○<χ|φ∧ψ>⌋"
  by (metis assms)
```

Figure 20.: Transitivity and totalness together (∃∀ rule,ct'd)

In Fig. 16, *Sledgehammer* shows the validity of the axioms of **E** holding independently of the properties assumed of the betterness relation. In Figs. 17 and 18, *Sledgehammer* confirms that the D* axiom and the Sp axiom call for totalnesss and transitivity, respectively. Similarly, Figs. 19 and 20 show that COK and CM call for *both* transitivity and totalness.

4.5. Discussion

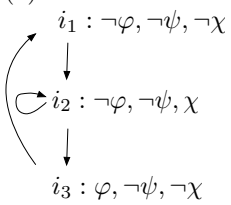
To conclude, with regards to correspondence, the situation for conditional (deontic) logic is still slightly different from the one for traditional modal logic. In the latter setting, the full equivalence between the property of the relation and the modal formula is verified by automated means. In the 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. van Benthem (2001), 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, cf. Benzmüller and Reiche (2022), an explicit dependency on the concrete 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

```
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))
```

(a) Model for HOL



(b) Preferential model. An arrow from i_1 to i_2 means $i_1 \succeq i_2$. No arrow from i_2 to i_1 means $i_2 \not\succeq i_1$

Figure 21.: A non-smooth model validating CM (max)

"CM $\Rightarrow$ m-smoothness" under the max rule (Fig. 10), *Nitpick* exhibits a model in which CM holds and m-smoothness falsified. This model is shown in Fig. 21. The corresponding preferential model is also shown below. 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 *Nitpick* always presents a finite standard model. We leave it as a topic for future research to investigate if the crucial distinction between standard and non-standard models for HOL which, according to Andrews (2002), 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 both the max rule and the opt rule neither reflexivity nor totalness correspond to an axiom. Finally, under the $\exists\forall$ 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.

5. The repugnant conclusion

In this section, we illustrate another possible use of the logical machinery. It consists in employing it for the computer-aided assessment of ethical arguments in philosophy. We use the repugnant conclusion discussed by Parfit (1984). We provide a computer encoding of his argument for the repugnant conclusion in order to make it amenable

to formal analysis and computer-assisted experiments. Isabelle/HOL is able to confirm the viability of a possible solution to the paradox, advocated by Temkin (1987) among others. It consists in resolving the paradox by abandoning the assumption of transitivity of "better than". Isabelle/HOL is also able to confirm the availability of a variant solution, which consists in weakening transitivity rather than abandonning it wholesale. Furthermore, one can show that not all the conceivable weakenings of transitivity will do.

The repugnant conclusion reads:

> "For any perfectly equal population with very high positive welfare, there is a population with very low positive welfare which is better, other things being equal" (Parfit, 1984, p. 523).

The target is "total utilitarianism", according to which the best outcome is given by the total of well-being in it (Parfit, 1984, p. 387). This view implies that any loss in the quality of lives in a population can be compensated for by a sufficient gain in the quantity of a population. Fig. 22 illustrates the repugnant conclusion. The blocks correspond to two populations, A and Z. The width of each block represents the number of people in the corresponding population, the height represents their quality of life. All the lives in the above diagram have lives worth living. People's quality of life is much lower in Z than in A but, since there are many more people in Z, there is a greater quantity of welfare in Z as compared to A. Consequently, although the people in A lead very good lives and the people in Z have lives only barely worth living, Z is nevertheless better than A according to classical utilitarianism.

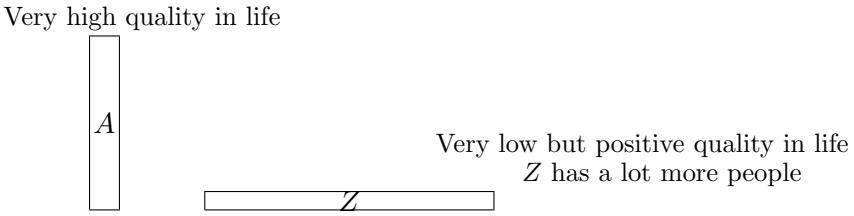


Figure 22.: Repugnant conclusion

It has been argued by e.g. Temkin (1987) that the repugnant conclusion can be blocked, by just dropping the assumption of the transitivity of "better than". This is best explained by considering a smaller version of the paradox, called the mere addition paradox. The repugnant conclusion is generated by iteration of the reasoning underlying the mere addition paradox.

The mere addition paradox is shown in Fig. 23. In population A, everybody enjoys a very high quality of life. In population A^+ there is one group of people as large as the group in A and with the same high quality of life. But A^+ also contains a number of people with a somewhat lower quality of life. In Parfit's terminology A^+ is generated from A by "mere addition". Population B has the same number of people as A^+ , their lives are worth living and at an average welfare level slightly above the average in A^+, but lower than the average in A. The link with the repugnant conclusion is that by reiterating this structure (scenario B^+ and C, C^+ etc.), we end up with a population Z in which all lives have a very low positive welfare.

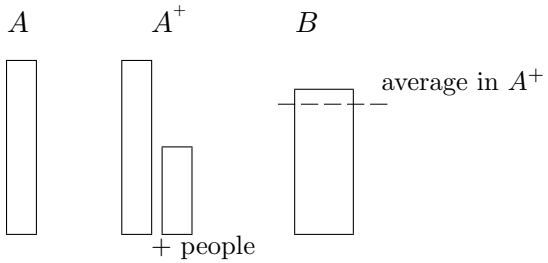


Figure 23.: Mere addition paradox

The following statements are all plausible:

(P0) A is strictly better than B: $A > B$. Otherwise, in the original scenario, by parity of reasoning or consistency (scenario B^+ and C, C^+ etc.) one would have to deny that A is better than Z.

(P1) A^+ is at least as good as A: $A^+ \geq A$. Justification: A^+ is not worse than (and hence at least as good as) A; the addition of lives worth living (the + people) cannot make a population worse.

(P2) B is strictly better than A^+: $B > A^+$. Justification: A^+ and B have the same size; the average welfare level in B is slightly above the average in A^+, and the distribution is uniform across members. So B is better in regard to both average welfare (and thus also total welfare) and equality.

The relations $\geq$ and $>$ appearing in (P0)-(P2) apply to propositional formulas. It is usual to take $\varphi \geq \psi$ as a shorthand of $P(\varphi/\varphi \vee \psi)$, and $\varphi > \psi$ as a shorthand of $\varphi \geq \psi$ and $\psi \not\geq \varphi$. (Cf. Lewis (1973)). This is shown in Table 3.

Definiendum	Definiens	Reading
$\varphi \geq \psi$	$P(\varphi/\varphi \vee \psi)$	φ permitted, if $\varphi \vee \psi$
$\varphi > \psi$	$P(\varphi/\varphi \vee \psi) \wedge \bigcirc(\neg\psi/\varphi \vee \psi)$	φ permitted and ψ forbidden, if $\varphi \vee \psi$

Table 3.: Preference on formulas

Fig. 24 shows the encoding of P0-P2 in terms of obligation.

```
6
7  consts a::σ aplus::σ b::σ
8
9  (*the mere addition scenario*)
10
11 axiomatization where
12 (* A is striclty better than B*)
13  P0: "⌊(¬⊙<¬a|aVb>∧⊙<¬b|aVb>)⌋" and
14 (* Aplus is at least as good as A*)
15  P1: "⌊¬⊙<¬aplus|aVaplus>⌋" and
16 (* B is strictly better than Aplus*)
17  P2: "⌊(¬⊙<¬b|aplusVb> ∧ ⊙<¬aplus|aplusVb>)⌋"
18
```

Figure 24.: Encoding of the mere addition scenario

Fig. 25 shows some sample queries run on the scenario. On l. 19, the assumption of the transitivity of the betterness relation (on possible worlds) is assumed. *Sledgehammer* shows the inconsistency of (P0)-(P2). On l. 23, the assumption of transitivity is

not assumed. *Nitpick* shows the satisfiability of (P0)-(P2). The model generated by *Nitpick* is shown in Fig. 26.

```
(* Sledgehammer finds P0-P2 inconsistent given
transitivity of the betterness relation in the models*)

theorem TransIncons:
  assumes transitivity
  shows False
  sledgehammer
  by (metis P0 P1 P2 assms)


(* Nitpick shows consistency in the absence of transitivity*)

lemma true
  nitpick [satisfy, card i=3]   (*model found*)
  oops
```

Figure 25.: Sample queries on (P0)-(P2)

```
Nitpick found a model for card i = 3:

  Constants:
    (r) =
      (λx. _)
      ((i₁, i₁) := True, (i₁, i₂) := True,
       (i₁, i₃) := False, (i₂, i₁) := False,
       (i₂, i₂) := True, (i₂, i₃) := True,
       (i₃, i₁) := True, (i₃, i₂) := True,
       (i₃, i₃) := True)
    a = (λx. _)(i₁ := False, i₂ := False, i₃ := True)
    aplus = (λx. _)(i₁ := False, i₂ := True, i₃ := False)
    b = (λx. _)(i₁ := True, i₂ := False, i₃ := False)
```

Figure 26.: A non-transitive model satisfying (P0)-(P2)

```
(* Nitpick shows consistency if transitivity is weakened into acyclicity or quasi-transitivity*)

theorem Acyclcons:
  assumes loopfree
  shows true
  nitpick [show_all,satisfy,card=3] (* model found for card i=3 *)
  oops

theorem Quasicons2:
  assumes Quasitransit
  shows true
  nitpick [show_all,satisfy,card=4]
  oops
```

Figure 27.: A-cyclicity and quasi-transitivity

```
(* Sledgehammer confirms inconsistency in the presence of the interval order condition*)

theorem io:
  assumes reflexivity Ferrers
  shows false
  sledgehammer
  by (metis P0 P1 P2 assms(2))
```

Figure 28.: Interval order

Nitpick is also able to confirm that the mere-addition paradox is avoided if transitivity is not rejected wholesale, but weakened into a-cyclicity or quasi-transitivity. This point has in general been overlooked in the literature. On the other hand, *Sledgehammer* can verify that this solution does not work for the interval order condition, which represents another candidate weakening of transitivity. The verifications are shown in Figs. 27 and 28.

6. Conclusion

Utilizing the LogiKEy methodology and framework we have developed mechanizations of extensions of Åqvist's preference-based system **E** for conditional obligation. We have illustrated the use of the resulting tool for (i) meta-logical studies and for (ii) object-level application studies in normative reasoning. Novel contributions, partly contributed by the automated reasoning tools in Isabelle/HOL, include the automated verification of the correspondence between semantic properties and modal axioms, and the formalization and mechanization of Parfit's argument for the repugnant conclusion. This one reveals the possibility of a take on the scenario usually under-appreciated in the literature, which consists in weakening transitivity rather than reject it wholesale. Future work includes the handling of the full equivalence between properties and formulas, the formalization of (and comparison with) other solutions to the repugnant conclusion, and the analysis of other variant paradoxes discussed in the literature.

Funding

The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme. Dr. X. Parent was funded in whole, or in part, by the Austrian Science Fund (FWF) [M3240 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.

References

Andrews, P. (2002). *An Introduction to Mathematical Logic and Type Theory*. Springer.

Benzmüller, C. (2019). Universal (meta-)logical reasoning: Recent successes. *Science of Computer Programming*, 172:48–62.

Benzmüller, C., Claus, M., and Sultana, N. (2015). Systematic verification of the modal logic cube in Isabelle/HOL. In Kaliszyk, C. and Paskevich, A., editors, *PxTP 2015*, volume 186, pages 27–41, Berlin, Germany. EPTCS.

Benzmüller, C., Farjami, A., and Parent, X. (2019). 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(5):733–755.

Benzmüller, C., Gabbay, D., Genovese, V., and Rispoli, D. (2012). Embedding and automating conditional logics in classical higher-order logic. *Annals of Mathematics and Artificial Intelligence*, 66(1-4):257–271.

Benzmüller, C., Parent, X., and van der Torre, L. (2020). Designing normative theories for ethical and legal reasoning: Logikey framework, methodology, and tool support. *Artificial Intelligence*, 287:103348.

Benzmüller, C. and Reiche, S. (2022). Automating public announcement logic with relativized common knowledge as a fragment of HOL in LogiKEy. *Journal of Logic and Computation*.

Blanchette, J. C., Böhme, S., and Paulson, L. C. (2013). Extending Sledgehammer with SMT solvers. *Journal of Automated Reasoning*, 51(1):109–128.

Blanchette, J. C., Kaliszyk, C., Paulson, L. C., and Urban, J. (2016). Hammering towards QED. *Journal of Formalized Reasoning*, 9(1):101–148.

Blanchette, J. C. and Nipkow, T. (2010). Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In Kaufmann, M. and Paulson, L. C., editors, *ITP 2010*, volume 6172 of *LNCS*, pages 131–146. Springer.

Chellas, B. (1975). Basic conditional logic. *Journal of Philosophical Logic*, 4(2):pp. 133–153.

Chellas, B. (1980). *Modal Logic*. Cambridge University Press, Cambridge.

Hansson, B. (1969). An analysis of some deontic logics. *Noûs*, 3(4):373–398.

Hughes, G. E. and Cresswell, M. J. (1984). *A companion to modal logic*. Methuen, London.

Lewis, D. (1973). *Counterfactuals*. Blackwell, Oxford.

Luce, R. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191.

Parent, X. (2014). Maximality vs. optimality in dyadic deontic logic. *Journal of Philosophical Logic*, 43(6):1101–1128.

Parent, X. (2015). Completeness of Åqvist's systems E and F. *Review of Symbolic Logic*, 8(1):164–177.

Parent, X. (2021). Preference semantics for dyadic deontic logic: a survey of results. In Gabbay, D., Horty, J., Parent, X., van der Meyden, R., and van der Torre, L., editors, *Handbook of Deontic Logic and Normative Systems*, pages 1–70. College Publications, London. UK. Volume 2.

Parent, X. (2022). On some weakened forms of transitivity in the logic of norms. In Giordanni, L. and Casini, G., editors, *NMR 2022: 20th International Workshop on Non-Monotonic Reasoning*. CEUR-WS. Extended abstract.

Parent, X. and van der Torre, L. (2021). *Introduction to Deontic Logic and Normative Systems*. College Publications, London.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Temkin, L. S. (1987). Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16(2):138–187.

Van Benthem, J. (2001). Correspondence theory. In Gabbay, D. M. and Guenthner, F., editors, *Handbook of Philosophical Logic*, pages 325–408. Springer Netherlands, Dordrecht.