

Conditional normative 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 June 25, 2024

ABSTRACT

We report on the mechanization of (preference-based) conditional normative 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 equivalence is automatically verified in one direction, leading from the property to the axiom. The second use is as a tool for assessing ethical arguments. We provide a computer encoding of a well-known paradox (or impossibility theorem) in population ethics, Parfit’s repugnant conclusion. While some have proposed overcoming the impossibility theorem by abandoning the presupposed transitivity of “better than,” our formalisation unveils a less extreme approach, suggesting among other things the option of weakening transitivity suitably rather than discarding it entirely. 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; Isabelle/HOL; correspondence; automated theorem proving; population ethics; mere addition/repugnant conclusion paradox; transitivity of betterness; weakenings of transitivity.

1. Introduction

We report on the mechanization of (preference-based) conditional normative 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 and colleagues. Secondly, SDL is vulnerable to the well-known deontic paradoxes, including in particular Chisholm’s paradox of contrary-to-duty obligation, see Parent

CONTACT X. Parent. Email: xavier.parent@tuwien.ac.at; C. Benzmüller: christoph.benzmueller@uni-bamberg.de

and van der Torre (2021) for details. We thus focus here on Dyadic Deontic Logics (DDLs) with a preference-based semantics, which originate from the works of Hansson (1969) and Lewis (1973). To represent conditional obligation sentences, an “intentional” dyadic operator (which is weaker than material implication) is employed. The proposed semantics generalizes that of SDL: the SDL-ish binary classification of states into good/bad is relaxed to allow for grades of ideality and accommodate classifications such as best, 2nd-best, and so forth. More specifically, 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. For further details on this framework, the reader is referred to the overview chapter by Parent (2021) found in the second volume of the *Handbook of Deontic Logic and Normative 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 for DDL. Benzmüller et al. (2015) write: “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, “correspondence” is taken in the same (broad) sense that Hughes and Cresswell have in mind when they 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 $K+\alpha$ 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)

This is different from correspondence theory in the sense of Sahlqvist (1975) and van Benthem (2001). Typically, Sahlqvist-style modal correspondence theory studies the equivalence between modal formulas and first-order formulas over Kripke frames via the so-called standard translation. The goal is to identify syntactic classes of modal formulas that can be shown to define first-order conditions on frames, and which are themselves computable via an algorithm. Correspondence theory in this sense has not been developed for preference-based dyadic deontic logic and conditional logic yet. This is in part due to the more complex form of the truth conditions for the conditional. The Sahlqvist/van Benthem method allows to establish an equivalence between an axiom and a property. By contrast, our method will give us only one direction of the equivalence, from the property to the axiom, but not yet the other direction.

A distinctive feature of our method is its flexibility. We will primarily deal with

¹For $i \succeq j$, read “ i is at least as good as j ”.

the conventional evaluation pattern in terms of best antecedent-worlds, distinguishing between two notions of best, optimality and maximality, as is the custom in rational choice theory. For the sake of completeness, we will also consider variant truth conditions that do not rely on the limit assumption, which some consider controversial. Notably, Lewis (1973) rejected this assumption. In a deontic context, it amounts to assuming the existence of “the best of all possible worlds”. For simplicity, we will confine our analysis to Lewis’s variant rule.

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 paradoxes or impossibility theorems in population ethics, the so-called “repugnant conclusion” due to Parfit (1984). We provide a computer encoding of the repugnant conclusion to make it amenable to formal analysis and computer-assisted experiments. We believe that the formalization has the potential to further stimulate the philosophical debate on the repugnant conclusion, given the considerable simplifications it achieves. Specifically, our formalization hints at the possibility of a fresh perspective on the scenario. While some have proposed overcoming the impossibility theorem by relinquishing the presupposed transitivity of “better than,” this solution is often deemed too radical. We distinguish between “better than” as a relation on formulas and as a relation on possible worlds, the second being used to elucidate the formal meaning of the first. Shifting the emphasis on the second, our formalisation unveils a less extreme approach. It consists in weakening transitivity suitably rather than discarding it entirely. However, we show that not all candidate weakenings of transitivity will do. In particular, drawing on Parent (2024), we argue that acyclicity (or even quasi-transitivity) does the job, but not the interval order condition. We also raise the question if transitivity is the sole cause of the paradox. We point out that under the standard interpretation of “best” in terms of maximality (quasi-)transitivity generates an inconsistency only if the set of possible worlds is assumed to be finite—an assumption that might appear overly limiting, if not arbitrary. This finding allows us to resolve (negatively) an open problem from previous work:² whether under the rule of maximality the finite model property holds for preference models with transitive, quasi-transitive, or interval order relations.

Until now, these particular points have remained unnoticed. Previously, one could verify them manually, but now, automation eliminates the need for logical expertise. Additionally, experimenting and implementing variations, such as changing the evaluation rule for the conditional, is straightforward. The practicality of the proposed tool lies in its ability to swiftly (dis-)confirm alternative hypotheses with minimal reliance on logical expertise. With this case study—the first of its kind—we hope to provide evidence that automated tools may help to facilitate the understanding and assessment of ethical arguments in philosophical debates. Previously, the second author utilized similar techniques in computational metaphysics. Notably, the inconsistency within the axioms of Gödel’s ontological argument went unnoticed until 2013, when it was automatically identified by the higher-order theorem prover Leo-II—see Benzmüller and Woltzenlogel Paleo (2016).

Readers should be warned that there is less standardization in preference semantics for dyadic deontic logic than in the usual Kripke-style semantics for (monadic) deontic logic, and more room for variation. This is because several factors must be juggled all at once. In this paper we stick to Åqvist (1987, 2002)’s approach, but the account is also applicable to further variants. Those who wish to get a general overview of the

²Cf. Goble (2019) and Parent (2021; 2024).

possible approaches that can be taken might find it useful to consult Makinson (1993). The interested reader will find in Goble (2019) and Parent (2021) further pointers to the literature.

The paper is organized as follows. Section 5 recalls system **E** and its extensions. Section 3 shows the embedding of **E** in Isabelle/HOL. Section 4 studies the correspondence between the properties of the betterness relation and the axioms. Section 5 discusses the repugnant conclusion. Section 6 concludes.³

2. System **E**

We describe the language, the semantics and proof theory of system **E** and its extensions.

2.1. Language

The language, call it \mathcal{L} , is defined by the following BNF:

$$\begin{aligned} \text{Atomic formulas: } & p \in \mathbb{P} \\ \text{Formulas: } & A \in \mathcal{L} \\ A ::= & p \mid \neg A \mid A \vee A \mid \Box A \mid \bigcirc(A/A) \end{aligned}$$

$\neg A$ is read as “not- A ”, and $A \vee B$ as “ A or B ”. $\Box A$ is read as “ A is settled as true”, and $\bigcirc(B/A)$ as “ B is obligatory, given A ”.

The Boolean connectives other than “ \neg ” and “ \vee ” are defined as usual. $\Diamond A$ is short for $\neg \Box \neg A$. $P(B/A)$ (“ B is permitted, given A ”) is short for $\neg \bigcirc(\neg B/A)$, $\bigcirc A$ (“ A is unconditionally obligatory”) and PA (“ A is unconditionally permitted”) are short for $\bigcirc(A/\top)$ and $P(A/\top)$, where \top denotes a tautology.

2.2. 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 (called its “universe”), \succeq is a preference relation ranking the elements of W in terms of betterness or comparative goodness, and v is a function assigning to each atomic formulas a subset of W (intuitively, the subset of those worlds where the atomic formula is true). $a \succeq b$ may be read “ a is at least as good as b ”. Also, \succ is the strict counterpart of \succeq , defined by $a \succ b$ (a is strictly better than b) iff $a \succeq b$ and $b \not\succeq a$. And \approx is the equal goodness relation, defined by $a \approx b$ (a and b are equally good) iff $a \succeq b$ and $b \succeq a$. For future reference, note that by definition \succ is irreflexive (for all a , $a \not\succeq a$) and asymmetric (for all a, b , if $a \succ b$ then $b \not\succeq a$).

A model M is said to be finite if its universe W is. The truth conditions for modal and deontic formulas read:

- $M, a \models \Box \varphi$ iff $\forall b \in W$ we have $M, b \models \varphi$

³The theory files are available for downloading at <http://logikey.org> under sub-repository “/Deontic-Logics/cube-dll/” (files “DDLcube.thy”, “mere.addition.opt.thy”, “mere.addition.max.thy” and “mere.addition.Lewis.thy”). A corresponding (but slightly modified) Isabelle/HOL dataset is presented in Parent and Benzmüller (2024).

- $M, a \models \bigcirc(\psi/\varphi)$ iff $\forall b \in \text{best}(\varphi)$ we have $M, b \models \psi$

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). A 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. An optimal element is maximal, but not the other way around. 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

$$a \in \max(\varphi) \Leftrightarrow a \models \varphi \ \& \ \neg \exists b (b \models \varphi \ \& \ b \succ a)$$

$$a \in \text{opt}(\varphi) \Leftrightarrow a \models \varphi \ \& \ \forall b (b \models \varphi \rightarrow a \succeq b)$$

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

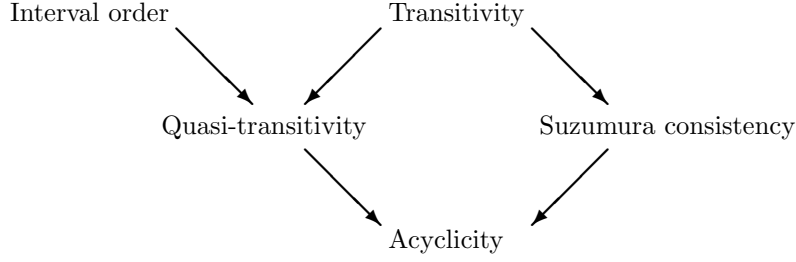
- Reflexivity: $a \succeq a$;
- Totality or (strong) connectedness: $a \succeq b$ or $b \succeq a$ (or both);
- Transitivity: if $a \succeq b$ and $b \succeq c$, then $a \succeq c$;
- Various weakenings of transitivity (from so-called rational choice theory):
 - Quasi-transitivity: if $a \succ b$ and $b \succ c$ then $a \succ c$;
 - Acyclicity: if $a \succ^* b$, then $b \not\succeq a$, where \succ^* is the transitive closure of \succ ;
 - Suzumura consistency: if $a \succeq^* b$, then $b \not\succeq a$, where \succeq^* is the transitive closure of \succeq ;
 - Interval order: \succeq is reflexive and Ferrers (if $a \succeq b$ and $c \succeq d$, then $a \succeq d$ or $c \succeq b$).

Intuitively, quasi-transitivity demands that the strict part of the betterness relation be transitive. Acyclicity rules out strict betterness cycles. Suzumura consistency rules out cycles with at least one instance of strict betterness. Acyclicity may be interpreted as generalizing asymmetry to a path of arbitrary length. Totality implies reflexivity. Given reflexivity and Ferrers, totality follows, and so the interval order condition can equivalently be defined by the pair “totality + Ferrers”. Intuitively, the interval order condition permits instances where transitivity of equal goodness fails, due to discrimination thresholds. These are cases where $a \approx b$ and $b \approx c$ but $a \not\approx c$ (see Luce 1956).

These weakenings of transitivity are discussed in greater depth in Parent (2024). Fig. 1 shows their relationships. An arrow from one condition to the other means that the first implies the second. The lack of an arrow between two conditions means that they are independent.⁴

⁴Proofs and additional discussion may be found in Parent (2024).

Figure 1.: Weakenings of transitivity (Parent (2024))



Lewis’s 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 $\text{best} \in \{\text{max}, \text{opt}\}$:

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.⁵ 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.

2.3. Systems

The relevant systems are shown in Fig. 2. A line between two systems indicates that the system to the left is strictly included in the system to the right. **E**, **F** and **G** are from Åqvist (1987). **F**+CM and **F**+DR are from Parent (2014) and Parent (2024), respectively.

⁵One 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)). Neighborhood semantics for dyadic deontic logic are investigated by Segerberg (1971), Nortmann (1986) and Goble (2004) among others.

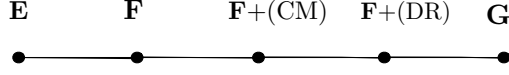


Figure 2.: Systems

All the systems contain the classical propositional calculus and the modal system S5.⁶ Then they add the following axiom schemata:

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

$$\begin{array}{ll}
 \text{S5-schemata for } \Box & (\text{S5}) \\
 \bigcirc(\psi \rightarrow \xi/\varphi) \rightarrow (\bigcirc(\psi/\varphi) \rightarrow \bigcirc(\xi/\varphi)) & (\text{COK}) \\
 \bigcirc(\psi/\varphi) \rightarrow \Box \bigcirc(\psi/\varphi) & (\text{Abs}) \\
 \Box\varphi \rightarrow \bigcirc(\varphi/\psi) & (\text{Nec}) \\
 \Box(\varphi \leftrightarrow \psi) \rightarrow (\bigcirc(\xi/\varphi) \leftrightarrow \bigcirc(\xi/\psi)) & (\text{Ext}) \\
 \bigcirc(\varphi/\varphi) & (\text{Id}) \\
 \bigcirc(\xi/\varphi \wedge \psi) \rightarrow \bigcirc(\psi \rightarrow \xi/\varphi) & (\text{Sh})
 \end{array}$$

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

$$\diamond\varphi \rightarrow (\bigcirc(\psi/\varphi) \rightarrow P(\psi/\varphi)) \quad (\text{D}^*)$$

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

$$(\bigcirc(\psi/\varphi) \wedge \bigcirc(\xi/\varphi)) \rightarrow \bigcirc(\xi/\varphi \wedge \psi) \quad (\text{CM})$$

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

$$\bigcirc(\xi/\varphi \vee \psi) \rightarrow (\bigcirc(\xi/\varphi) \vee \bigcirc(\xi/\psi)) \quad (\text{DR})$$

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

$$(P(\psi/\varphi) \wedge \bigcirc(\psi \rightarrow \xi/\varphi)) \rightarrow \bigcirc(\xi/\varphi \wedge \psi) \quad (\text{Sp})$$

We give an intuitive explanation for these axioms. COK is the conditional analog 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 dyadic 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 analog of the identity principle. Sh is named after (Shoham, 1988, p. 77), who seems to have been the first to discuss it. One can see it as the deontic analog of one-half of the deduction theorem. D^{*} is the conditional analog of the familiar D axiom. In its equivalent form, $\diamond\varphi \rightarrow \neg(\bigcirc(\psi/\varphi) \wedge \bigcirc(\neg\psi/\varphi))$, this axiom rules out the possibility of conflicts between obligations arising in a context φ that is possible. CM and DR correspond to the principle of cautious monotony and disjunctive rationality from the non-monotonic logic literature –Kraus et al. (1990). 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 (1975), Sp is best explained using the (more widely known) principle of rational monotony RM from non-monotonic

⁶S5 is characterized by the rule of necessitation (“If $\vdash A$, then $\vdash \Box A$ ”), and the K, T and 5 axioms (5 is $\diamond A \rightarrow \Box \diamond A$).

logic—see Kraus et al. (1990). The two laws are inter-derivable in **E**. RM is obtained by replacing, in Sp, $\bigcirc(\psi \rightarrow \xi/\varphi)$ with $\bigcirc(\xi/\varphi)$, to read:

$$(P(\psi/\varphi) \wedge \bigcirc(\xi/\varphi)) \rightarrow \bigcirc(\xi/\varphi \wedge \psi) \quad (\text{RM})$$

RM says that realizing a permission does not modify our 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, Parent (2021; 2024)). *(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.*

In part (i), (ii) and (iii) of Th. 2.1, and in Th. 2.2, soundness and completeness are taken in their strong sense. They establish a correspondence between the syntactic and semantic consequence relation while also accommodating a potentially infinite set of assumptions. To be more precise, the theorems are of the form: where Γ is a set of formulas or assumptions, $\Gamma \vdash \varphi$ if and only if $\Gamma \models \varphi$. In part (iv) of Th. 2.1, soundness and completeness are taken in their weak sense: Γ is required to be finite; this amounts to establishing a match between theorems and validities only. This restriction is because the completeness proof appeals in an essential way to the assumption that models are finite—for more details, see (Parent, 2024, §4).

Theorem 2.2 (Soundness and completeness, Parent (2014)). *(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, and additional results, we refer the reader to Parent (2021; 2024).

2.4. Correspondences

Table 1 shows some of the known “correspondences” between semantic properties and formulas as extracted from Th. 2.1 and Th. 2.2. Thus, the term “correspondence” is understood along the lines suggested by Hughes and Cresswell (Cf. Section 1). 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. 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.

To improve readability, we have used certain shortcuts, albeit with the potential drawback of simplifying the data. The lack of correspondence in the 1st, 2nd and 5th

Property	Axiom (max)	Axiom (opt)	Ref.
reflexivity	×	×	Parent (2015)
totality	×	×	Parent (2015)
limitedness	D*	D*	Parent (2015)
smoothness	CM	CM	Parent (2014)
transitivity	×	Sp	Parent (2014; 2024)
transitivity+ totality	Sp	×	Parent (2014)
interval order	DR	DR	Parent (2024)

Table 1.: Some correspondences

row (starting from the top, going downwards) is for the general case, when no constraint is put on \succeq . Thus, assuming one of reflexivity, totality or transitivity (under the max) does not add new validities. Similarly, the correspondence for limitedness is independent of any other properties (or axioms) in the background. The correspondence results for smoothness, transitivity (under the opt), transitivity+totality (under the max) and interval order assume D* and limitedness in the background. Quasi-transitivity, Suzumura consistency and acyclicity are known not to correspond to any formula in the general case, under the max rule—Parent (2024). This holds whether or not limitedness or smoothness is assumed in the background. However, it is not known what happens under the opt rule. Therefore, we have put these three conditions aside.

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 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

```

1 theory DDLcube imports Main (** Benzmueller/Parent 2024 **)
2
3 begin (* Settings for model finder Nitpick *)
4
5 nitpick_params [user_axioms, show_all, expect=genuine, format=2]
6
7 (** We introduce Agvist's system E from the 2019 IfColog paper **)
8
9 typedecl i (* Possible worlds *)
10 type_synonym  $\sigma$  = "(i $\Rightarrow$ bool)"
11 type_synonym  $\alpha$  = "i $\Rightarrow$  $\sigma$ " (* Type of betterness relation between worlds *)
12 type_synonym  $\tau$  = " $\sigma\Rightarrow\sigma$ "
13
14 consts aw::i (* Actual world *)
15 abbreviation etrue :: " $\sigma$ " ("T") where "T  $\equiv$   $\lambda w$ . True"
16 abbreviation efalse :: " $\sigma$ " ("⊥") where "⊥  $\equiv$   $\lambda w$ . False"
17 abbreviation enot :: " $\sigma\Rightarrow\sigma$ " ("¬"[52]53) where "¬ $\varphi$   $\equiv$   $\lambda w$ .  $\neg\varphi(w)$ "
18 abbreviation eand :: " $\sigma\Rightarrow\sigma\Rightarrow\sigma$ " (infixr"∧"51) where " $\varphi\wedge\psi$   $\equiv$   $\lambda w$ .  $\varphi(w)\wedge\psi(w)$ "
19 abbreviation eor :: " $\sigma\Rightarrow\sigma\Rightarrow\sigma$ " (infixr"∨"50) where " $\varphi\vee\psi$   $\equiv$   $\lambda w$ .  $\varphi(w)\vee\psi(w)$ "
20 abbreviation eimpf :: " $\sigma\Rightarrow\sigma\Rightarrow\sigma$ " (infixr"→"49) where " $\varphi\rightarrow\psi$   $\equiv$   $\lambda w$ .  $\varphi(w)\rightarrow\psi(w)$ "
21 abbreviation eimpb :: " $\sigma\Rightarrow\sigma\Rightarrow\sigma$ " (infixr"←"49) where " $\varphi\leftarrow\psi$   $\equiv$   $\lambda w$ .  $\psi(w)\rightarrow\varphi(w)$ "
22 abbreviation eequ :: " $\sigma\Rightarrow\sigma\Rightarrow\sigma$ " (infixr"↔"48) where " $\varphi\leftrightarrow\psi$   $\equiv$   $\lambda w$ .  $\varphi(w)\leftrightarrow\psi(w)$ "
23
24 abbreviation ebox :: " $\sigma\Rightarrow\sigma$ " ("□") where "□ $\varphi$   $\equiv$   $\lambda w$ .  $\forall v$ .  $\varphi(v)$ "
25 abbreviation ddediamond :: " $\sigma\Rightarrow\sigma$ " ("◇") where "◇ $\varphi$   $\equiv$   $\lambda w$ .  $\exists v$ .  $\varphi(v)$ "
26
27 abbreviation evalid :: " $\sigma\Rightarrow$ bool" ("□"[8]109) (* Global validity *)
28   where "[p]  $\equiv$   $\forall w$ . p w"
29 abbreviation ecjactual :: " $\sigma\Rightarrow$ bool" ("□i"[7]105) (* Local validity — in world aw *)
30   where "[p]i  $\equiv$  p(aw)"
31
32 consts r :: " $\alpha$ " (infixr "r" 70) (* Betterness relation *)

```

Figure 3.: Basic semantical ingredients; propositional and modal connectives

```

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

```

Figure 4.: Truth conditions

SMT (satisfiability modulo theories) solvers, and many of these systems in turn use efficient SAT solver technology internally. Proof automation with *Sledgehammer* and (counter-)model finding with *Nitpick* were invaluable in supporting our exploratory modeling approach. These tools were very responsive in automatically proving (*Sledgehammer*), disproving (*Nitpick*), or showing consistency by providing a model (*Nitpick*). 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**

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).

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

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. 3 shows the basic ingredients in the preference 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. 32). In Fig. 4, 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. 55) and to verify the non-equivalence between the two types of truth conditions (l. 61). *Sledgehammer* is able to show the validity of all the axioms of **E**. This is shown in Fig. 5 for the max rule. It takes only a few ms for some provers to prove a formula. For instance *cvc4* shows Abs in 1ms, and Sh in 10 ms.

```

184 (*****
185 Tests max rule
186 *****)
187
188 (*inference rules*)
189
190 lemma MP: "[[ $\varphi$ ]; [ $\varphi \rightarrow \psi$ ]]  $\implies$  [ $\psi$ ]" by simp
191 lemma Nec: "[ $\varphi$ ]  $\implies$  [ $\Box\varphi$ ]" by simp
192
193 (* " $\Box$ " is an S5 modality *)
194 lemma C_1_refl: "[ $\Box\varphi \rightarrow \varphi$ ]" by simp
195 lemma C_1_trans: "[ $\Box\varphi \rightarrow (\Box(\Box\varphi))$ ]" by simp
196 lemma C_1_sym: "[ $\varphi \rightarrow (\Box(\Diamond\varphi))$ ]" by simp
197
198 (* Axioms of E holds *)
199
200 lemma Abs: "[ $\Box\langle\psi|\varphi\rangle \rightarrow \Box\langle\psi|\varphi\rangle$ ]" sledgehammer oops
201 lemma Nec: "[ $\Box\psi \rightarrow \Box\langle\psi|\varphi\rangle$ ]" sledgehammer oops
202 lemma Ext: "[ $\Box(\varphi_1 \leftrightarrow \varphi_2) \rightarrow (\Box\langle\psi|\varphi_1\rangle \leftrightarrow \Box\langle\psi|\varphi_2\rangle)$ ]" sledgehammer oops
203 lemma Id: "[ $\Box\langle\varphi|\varphi\rangle$ ]" sledgehammer oops
204 lemma Sh: "[ $\Box\langle\psi|\varphi_1 \wedge \varphi_2\rangle \rightarrow \Box(\langle\varphi_2 \rightarrow \psi\rangle | \varphi_1\rangle)$ ]" sledgehammer oops
205 lemma COK: "[ $\Box\langle\psi_1 \rightarrow \psi_2\rangle | \varphi\rangle \rightarrow (\Box\langle\psi_1|\varphi\rangle \rightarrow \Box\langle\psi_2|\varphi\rangle)$ ]" sledgehammer oops

```

Figure 5.: Axioms of **E** (max)

3.3. Properties

The encoding of the properties of the betterness relation are shown in Figs. 6 and 7. On l. 99-104 of Fig. 6, one sees the different versions of Lewis' limit assumption. The property in Fig. 7 is the interval order condition. This one is usually described as the combination of totality with the Ferrers condition encoded in l. 146. *Sledgehammer* confirms a fact often overlooked in the literature, that totality can be replaced by the simpler condition of reflexivity (l. 149-152). The other candidate weakenings of transitivity discussed earlier are also encoded in the theory file. For simplicity's sake, we only give the example of quasi-transitivity and acyclicity. The encoding of the second is shown in Fig. 8. In Isabelle/HOL, the transitive closure of a relation can be defined in a few lines, shown in Fig. 9. The encoding of quasi-transitivity is shown in Fig. 10.

4. Verifying the correspondences

In this section, the correspondences for the axioms are investigated. The task is to automatically verify that a given property is sufficient for the validation of the corre-

```

87 (*****
88 Properties
89 *****)
90
91 (* The standard properties of the betterness relation *)
92
93 abbreviation "reflexivity ≡ (∀x. x r x)"
94 abbreviation "transitivity ≡ (∀x y z. (x r y ∧ y r z) → x r z)"
95 abbreviation "totalness ≡ (∀x y. (x r y ∨ y r x))"
96
97 (* 4 versions of Lewis's limit assumption *)
98
99 abbreviation "mlimitedness ≡ (∀φ. (∃x. (φ)x) → (∃x. max<φ>x))"
100 abbreviation "msmoothness ≡
101 (∀φ x. ((φ)x → (max<φ>x ∨ (∃y. (y r x ∧ ¬(x r y) ∧ max<φ>y))))))"
102 abbreviation "olimitedness ≡ (∀φ. (∃x. (φ)x) → (∃x. opt<φ>x))"
103 abbreviation "osmoothness ≡
104 (∀φ x. ((φ)x → (opt<φ>x ∨ (∃y. (y r x ∧ ¬(x r y) ∧ opt<φ>y))))))"

```

Figure 6.: Standard properties

```

144 (* Interval order (reflexivity + Ferrers) *)
145
146 abbreviation Ferrers
147 where "Ferrers ≡ (∀x y z u. (x r u ∧ y r z) → (x r z ∨ y r u))"
148
149 theorem T2:
150 assumes Ferrers and reflexivity (*fact overlooked in the literature*)
151 shows totality
152 by (simp add: assms(1) assms(2)) (* proof found *)

```

Figure 7.: Interval order

sponding axiom as per Table 1. We begin by assuming that the truth conditions for the obligation operator are given in terms of maximality, go on to consider the case where they are given in terms of optimality, and finally extend the scope of our inquiry to a well-known variant evaluation rule for the conditional due to Lewis (1973). The three evaluation rules collapse only in the presence of all the properties of the betterness relation, including limitedness (which famously Lewis rejected). The consideration of Lewis’s evaluation rule will also be needed for the case study in Section 5.

4.1. Max rule

Here we check known correspondences for the max rule. *Sledgehammer* and *Nitpick* confirm that an axiom is not valid unless the matching property is assumed:

- If the relevant property is not assumed, counter-models for the corresponding axiom (D*, CM, DR and Sp) are found by *Nitpick*. This is Figs. 11, 13 and 14;
- If the property is assumed, then the corresponding axiom is proved by *Sledgehammer*. Fig. 12 shows it for limitedness and smoothness, Fig. 13 for the interval order condition, and Fig. 14 for the combination of transitivity and totality.

The implications having the form “property \Rightarrow axiom” are all verified. However, the converse implications are all falsified by *Nitpick*. We will come back to this point later on.

```

139 (* A-cyclicity: cycles of strict betterness are ruled out*)
140
141 abbreviation loopfree
142   where "loopfree  $\equiv \forall x y. \text{tcr\_strict } x y \longrightarrow (y \text{ r } x \longrightarrow x \text{ r } y)"$ "

```

Figure 8.: Acyclicity

```

106 (* Weaker forms of transitivity--they require the notion of
107 transitive closure*)
108
109 definition transitive :: " $\alpha \Rightarrow \text{bool}$ "
110   where "transitive Rel  $\equiv \forall x y z. \text{Rel } x y \wedge \text{Rel } y z \longrightarrow \text{Rel } x z"$ "
111 definition sub_rel :: " $\alpha \Rightarrow \alpha \Rightarrow \text{bool}$ "
112   where "sub_rel Rel1 Rel2  $\equiv \forall u v. \text{Rel1 } u v \longrightarrow \text{Rel2 } u v"$ "
113 definition assfactor :: " $\alpha \Rightarrow \alpha$ "
114   where "assfactor Rel  $\equiv \lambda u v. \text{Rel } u v \wedge \neg \text{Rel } v u "$ "
115
116 (* In HOL the transitive closure of a relation can be defined in a single line;
117 here we apply the construction to betterness relation r and for its strict
118 variant ( $\lambda u v. u \text{ r } v \wedge \neg v \text{ r } u$ ) *)
119 definition tcr
120   where "tcr  $\equiv \lambda x y. \forall Q. \text{transitive } Q \longrightarrow (\text{sub\_rel } r \ Q \longrightarrow Q \ x \ y)"$ "
121
122 definition tcr_strict
123   where "tcr_strict  $\equiv \lambda x y. \forall Q. \text{transitive } Q$ 
124          $\longrightarrow (\text{sub\_rel } (\lambda u v. u \text{ r } v \wedge \neg v \text{ r } u) \ Q \longrightarrow Q \ x \ y)"$ "

```

Figure 9.: Transitive closure

4.2. Opt rule

The outcome of our experiment is the same as for the max rule except for one small change. Transitivity no longer needs totality 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 formulas, defined by $\varphi \geq \psi$ iff $P(\varphi/\varphi \vee \psi)$. This is shown in Fig. 16. Again the converse implication is falsified by *Nitpick* (l. 356-361).

4.3. Inclusion

In the work of Benzmlüller 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.

```

126 (* Quasi-transitivity: the strict betterness
127 relation is transitive*)
128 abbreviation Quasitransit
129   where "Quasitransit  $\equiv \forall x y z. (\text{assfactor } r x y \wedge$ 
130          $\text{assfactor } r y z) \longrightarrow \text{assfactor } r x z$ "

```

Figure 10.: Quasi-transitivity

```

211 (* Max-Limitedness corresponds to D *)
212
213 lemma "[ $\diamond\varphi \rightarrow (\bigcirc\langle\psi|\varphi\rangle \rightarrow \text{P}\langle\psi|\varphi\rangle)$ ]"
214 nitpick [card i=3] (* counterexample found for card i=3 *)
215 oops
216
217 lemma "[ $(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\chi|\varphi\rangle) \rightarrow \bigcirc\langle\chi|\varphi\wedge\psi\rangle$ ]"
218 nitpick [card i=3] (* counterexample found *)
219 oops
220
221 lemma "[ $\bigcirc\langle\chi|(\varphi\vee\psi)\rangle \rightarrow ((\bigcirc\langle\chi|\varphi\rangle) \vee (\bigcirc\langle\chi|\psi\rangle))$ ]"
222 nitpick [card i=3] (* counterexample found *)
223 oops

```

Figure 11.: D*, CM and DR invalid in general

4.4. The $\exists\forall$ truth conditions (Lewis)

We extend the scope of our inquiry to other truth conditions for the conditional. Here we consider the variant rule proposed by Lewis (1973). In order to avoid any commitment to the limit assumption, he 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 $\varphi \rightarrow \psi$ -worlds, in which the obligation is never violated. Formally:

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

We shall refer to the statement appearing at the right-hand-side of “iff” as the $\exists\forall$ rule. The encoding is shown in Fig.17.

Isabelle/HOL can 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. 18. On l. 394, the validity of DEX is established under the max rule. On l. 398, DEX 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. 18 too. T18 shows the equivalence between the $\exists\forall$ rule and the opt rule, and T19 shows the equivalence between the $\exists\forall$ 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 avail-

```

224 theorem T8:
225   assumes mlimitedness
226   shows "D*": "[ $\Diamond\varphi \rightarrow \bigcirc\langle\psi|\varphi\rangle \rightarrow P\langle\psi|\varphi\rangle$ ]"
227   sledgehammer
228   by (metis assms)
229
230 lemma
231   assumes "D*": "[ $\Diamond\varphi \rightarrow \neg(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\neg\psi|\varphi\rangle)$ ]"
232   shows mlimitedness
233   nitpick [card i=3] (* counterexample found *)
234   oops
235
236 (* Smoothness corresponds to cautious monotony *)
237
238 theorem T9:
239   assumes msmoothness
240   shows CM: "[ $(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\chi|\varphi\rangle) \rightarrow \bigcirc\langle\chi|\varphi\wedge\psi\rangle$ ]"
241   sledgehammer
242   using assms by force
243
244 lemma
245   assumes CM: "[ $(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\chi|\varphi\rangle) \rightarrow \bigcirc\langle\chi|\varphi\wedge\psi\rangle$ ]"
246   shows msmoothness
247   nitpick [card i=3] (* counterexample found *)
248   oops
249

```

Figure 12.: Limit assumption

```

250 (*Interval order corresponds to disjunctive rationality*)
251
252 lemma
253   assumes reflexivity
254   shows DR: "[ $\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)$ ]"
255   nitpick [card i=3] (* counterexample found *)
256   oops
257
258 theorem T10:
259   assumes reflexivity and Ferrers
260   shows DR: "[ $\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)$ ]"
261   sledgehammer
262   by (metis assms(1) assms(2))
263
264 lemma
265   assumes DR: "[ $\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)$ ]"
266   shows reflexivity
267   nitpick [card i=1] (* counterexample found *)
268   oops
269
270 lemma
271   assumes DR: "[ $\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)$ ]"
272   shows Ferrers
273   nitpick [card i=2] (* counterexample found *)
274   oops

```

Figure 13.: Interval order

able for the strongest system \mathbf{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


```

276 (*Transitivity and totalness corresponds to the Spohn axiom (Sp)*)
277
278 Lemma
279   assumes transitivity
280   shows Sp: "[ ( P< $\psi$ | $\varphi$ >  $\wedge$  O<( $\psi$  $\rightarrow$  $\chi$ )| $\varphi$ > )  $\rightarrow$  O< $\chi$ |( $\varphi$  $\wedge$  $\psi$ )>]"
281   nitpick [card i=3] (* counterexample found *)
282   oops
283
284 Lemma
285   assumes totality
286   shows Sp: "[ ( P< $\psi$ | $\varphi$ >  $\wedge$  O<( $\psi$  $\rightarrow$  $\chi$ )| $\varphi$ > )  $\rightarrow$  O< $\chi$ |( $\varphi$  $\wedge$  $\psi$ )>]"
287   nitpick [card i=3] (* counterexample *)
288   oops
289
290 theorem T11:
291   assumes transitivity and totality
292   shows Sp: "[ ( P< $\psi$ | $\varphi$ >  $\wedge$  O<( $\psi$  $\rightarrow$  $\chi$ )| $\varphi$ > )  $\rightarrow$  O< $\chi$ |( $\varphi$  $\wedge$  $\psi$ )>]"
293   by (metis assms)

```

Figure 14.: Transitivity and totality (max)

```

295 theorem T12:
296   assumes transitivity and totality
297   shows transit: "[ ( P< $\varphi$ | $\varphi$  $\vee$  $\psi$ >  $\wedge$  P< $\psi$ | $\psi$  $\vee$  $\chi$ > )  $\rightarrow$  P< $\varphi$ |( $\varphi$  $\vee$  $\chi$ )>]"
298   by (metis assms(1) assms(2))
299
300 Lemma
301   assumes Sp: "[ ( P< $\psi$ | $\varphi$ >  $\wedge$  O<( $\psi$  $\rightarrow$  $\chi$ )| $\varphi$ > )  $\rightarrow$  O< $\chi$ |( $\varphi$  $\wedge$  $\psi$ )>]"
302   shows totality
303   nitpick [card i=1] (* counterexample found *)
304   oops
305
306 Lemma
307   assumes Sp: "[ ( P< $\psi$ | $\varphi$ >  $\wedge$  O<( $\psi$  $\rightarrow$  $\chi$ )| $\varphi$ > )  $\rightarrow$  O< $\chi$ |( $\varphi$  $\wedge$  $\psi$ )>]"
308   shows transitivity
309   nitpick [card i=2] (* counterexample found *)

```

Figure 15.: Transitivity and totality (max, cont.)

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 \mathbf{G}	Property (or pair of properties) of \succeq
(D *)	totality
(Sp)	transitivity
(COK)	transitivity and totality
(CM)	transitivity and totality

Table 2.: Axioms and properties under the $\exists\forall$ rule—from Parent (2021)

In Fig. 19, *Sledgehammer* shows the validity of the axioms of \mathbf{E} holding independently of the properties assumed of the betterness relation. In Figs. 20 and 21, *Sledgehammer* confirms that the D * axiom and the Sp axiom call for totality and transitivity, respectively. Similarly, Fig. 22 shows that COK and CM call for *both* transitivity and

```

344 (*transitivity*)
345
346 theorem T15:
347   assumes transitivity
348   shows Sp': "[ ( P<ψ|φ> ∧ ◊<(ψ→χ)|φ> ) → ◊<χ|(φ∧ψ)> ]"
349   by (metis assms)
350
351 theorem T16:
352   assumes transitivity
353   shows Trans': "[ ( P<φ|φ∨ψ> ∧ P<ψ|ψ∨ξ> ) → P<φ|φ∨ξ> ]"
354   by (metis assms)
355
356 lemma
357   assumes Sp: "[ ( P<ψ|φ> ∧ ◊<(ψ→χ)|φ> ) → ◊<χ|(φ∧ψ)> ]"
358   assumes Trans: "[ ( P<φ|φ∨ψ> ∧ P<ψ|ψ∨ξ> ) → P<φ|φ∨ξ> ]"
359   shows transitivity
360   nitpick [card i=2] (* counterexample found *)
361   oops

```

Figure 16.: Transitivity (opt)

```

65 (* David Lewis's evaluation rule for the conditional *)
66
67 abbreviation lewcond :: "σ⇒σ⇒σ" ("◊<_|>")
68   where "◊<ψ|φ> ≡ λv. (¬(∃x. (φ)(x)) ∨
69     (∃x. ((φ)(x) ∧ (ψ)(x) ∧ (∀y. ((y r x) → (φ)(y) → (ψ)(y))))))"
70 abbreviation lewperm :: "σ⇒σ⇒σ" ("f<_|>")
71   where "f<ψ|φ> ≡ ¬◊<¬ψ|φ>"

```

Figure 17.: $\exists\forall$ rule

totality. In all these cases, *Sledgehammer* fails to show the converse implications.

4.5. Assessment

As mentioned in Sect. 3.1, we found that the tools are very responsive in automatically proving (*Sledgehammer*), disproving (*Nitpick*), or showing consistency by providing a model (*Nitpick*). Not only an answer is returned, but also a justification for this answer. However, concerning this, we found that *Nitpick* fares better than *Sledgehammer*. When *Sledgehammer* has proved a theorem successfully, the list of the definitions, axioms and lemmas to be used is returned. But the derivation itself is not given.⁷ In principle one could look into the detailed proof output file of the external provers called by *Sledgehammer*, but this requires technical expertise. A simple example is given in Fig. 23. Line 161 tells us that quasi-transitivity follows from transitivity by assumption (“assms”) and using the definition of an asymmetric factor (“assfactor”). This is indeed how this would be shown by hand. However, a detailed argument is not given. By contrast, *Nitpick* always gives the full details of the model justifying its answer, and this one was always correct in our experiments. An example of such a model is given in Fig. 24, which we will discuss in a moment.

A comparative study with native provers, similar to the one in Steen et al. (2023), must be left as a topic for future work. We are not aware of a similar automation of

⁷We tried to use Isar style reconstruction of proofs with *Sledgehammer*, but without much success for our examples.

```

389 (*****
390 Relationship Lewis rule and max/opt rule
391 *****)
392
393 (* deontic explosion-max rule *)
394 theorem DEX: "[( $\Diamond\varphi \wedge \circ\langle\psi|\varphi\rangle \wedge \circ\langle\neg\psi|\varphi\rangle$ )  $\rightarrow \circ\langle\chi|\varphi\rangle$ ]"
395 by blast
396
397 (* no-deontic explosion-lewis rule *)
398 lemma DEX: "[( $\Diamond\varphi \wedge \circ\langle\psi|\varphi\rangle \wedge \circ\langle\neg\psi|\varphi\rangle$ )  $\rightarrow \circ\langle\chi|\varphi\rangle$ ]"
399 nitpick [card i=2] (* counterexample found*)
400 oops
401
402 theorem T18:
403   assumes mlimitedness and transitivity and totality
404   shows " $\circ\langle\psi|\varphi\rangle \leftrightarrow \circ\langle\psi|\varphi\rangle$ "
405   sledgehammer
406   by (smt (z3) assms)
407
408 theorem T19:
409   assumes mlimitedness and transitivity and totality
410   shows " $\circ\langle\psi|\varphi\rangle \leftrightarrow \circ\langle\psi|\varphi\rangle$ "
411   sledgehammer
412   by (smt (z3) assms)

```

Figure 18.: Deontic explosion (DEX)

```

415 (***)
416 axioms of E holding irrespective of the properties of r
417 *****)
418
419 theorem Abs: "[ $\circ\langle\psi|\varphi\rangle \rightarrow \Box\circ\langle\psi|\varphi\rangle$ ]" (*sledgehammer*) oops
420 theorem Nec: "[ $\Box\psi \rightarrow \circ\langle\psi|\varphi\rangle$ ]" (*sledgehammer*) oops
421 theorem Ext: "[ $\Box(\varphi_1 \leftrightarrow \varphi_2) \rightarrow (\circ\langle\psi|\varphi_1\rangle \leftrightarrow \circ\langle\psi|\varphi_2\rangle)$ ]" sledgehammer oops
422 theorem Id: "[ $\circ\langle\varphi|\varphi\rangle$ ]" (*sledgehammer*) oops
423 theorem Sh: "[ $\circ\langle\psi|\varphi_1 \wedge \varphi_2\rangle \rightarrow \circ\langle\varphi_2 \rightarrow \psi|\varphi_1\rangle$ ]" (*sledgehammer*) oops

```

Figure 19.: Axioms independent of the properties ($\exists\forall$ rule)

the systems studied in this paper using other methods.⁸ A comparison with a prover for a related system (e.g. KLMLean 2.0, due to Giordano et al. (2007)) would already be beneficial.

The entire Isabelle document (“DDLcube.thy”) is verified by Isabelle2023 in 1m50s on an Apple M1 with 8 GB of memory. During this time, Isabelle/HOL solves 82 problems, whereby demonstrating good responsiveness. It takes 15s for *Nitpick* to find 40 (counter-)models, and 1m35s for *Sledgehammer* to show the validity of 34 formulas and verify 6 implication relations between properties of the betterness relation. Additionally, we consistently observe accurate proofs and models, contrasting with the inherently error-prone nature of the traditional pen-and-paper method. The assurance of accuracy is an added benefit of the known faithfulness of the embedding, Th. 3.1, a distinctive hallmark of the method.

We end with a critical assessment of the findings on correspondence. The situation for conditional (deontic) logic is still slightly different from the one for traditional

⁸The hyper-sequent system for **E** defined in Ciabattoni et al. (2022) comes with a method for extracting a counter-model from a failed proof search. It holds the promise to provide a theorem prover against which we could evaluate the one described in this paper.

```

449 (* axioms of E holding if r total *)
450
451 theorem T20:
452   assumes totality
453   shows "[ $\Diamond\psi \rightarrow (\circ\langle\psi|\varphi\rangle \rightarrow f\langle\psi|\varphi\rangle)$ ]"
454   by (metis assms)

```

Figure 20.: Totality alone ($\exists\forall$ rule)

```

465 theorem T21:
466   assumes transitivity
467   shows Sp'': "[ $(f\langle\psi|\varphi\rangle \wedge \circ\langle\psi\rightarrow\chi|\varphi\rangle) \rightarrow \circ\langle\chi|(\varphi\wedge\psi)\rangle$ ]"
468   (*sledgehammer*)
469   using assms by blast
470
471 theorem T22:
472   assumes transitivity
473   shows Tr'': "[ $(f\langle\varphi|\varphi\vee\psi\rangle \wedge f\langle\psi|\psi\vee\chi\rangle) \rightarrow f\langle\varphi|\varphi\vee\chi\rangle$ ]"
474   (*sledgehammer*)
475   using assms by blast
476
477 lemma
478   assumes Sp'': "[ $(f\langle\psi|\varphi\rangle \wedge \circ\ddagger(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \circ\langle\chi|(\varphi\wedge\psi)\rangle$ ]"
479   shows transitivity
480   nitpick (* counterexample found *)
481   oops
482
483 lemma
484   assumes Tr'': "[ $(f\langle\varphi|\varphi\vee\psi\rangle \wedge f\langle\psi|\psi\vee\chi\rangle) \rightarrow f\langle\varphi|\varphi\vee\chi\rangle$ ]"
485   shows transitivity
486   nitpick (* counterexample found *)
487   oops

```

Figure 21.: Transitivity alone ($\exists\forall$ rule, ct'd)

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 atomic formulas 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 atomic formulas. We have worked at the level of models. But in so-called correspondence theory *à la* Salqvist-van Benthem, 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. Benz Müller and Reiche (2022), an explicit dependency on the concrete evaluation domain has been modeled. The question as to whether this idea can be extended to support a notion of frame-validity is a topic for future research.

Second, the most we got is that a given property is a sufficient condition for the

```

500 (* axioms of G holding if r both transitive and total *)
501
502 theorem T23:
503   assumes transitivity and totality
504   shows COK: "[o<(\psi_1 \to \psi_2) | \varphi> \to (o<\psi_1 | \varphi> \to o<\psi_2 | \varphi>)]"
505   by (smt (verit, ccfv_SIG) assms(1) assms(2))
506
507 theorem T24:
508   assumes transitivity and totality
509   shows CM'': "[ (o<\psi | \varphi> \wedge o<\chi | \varphi>) \to o<\chi | \varphi \wedge \psi> ]"
510   by (metis assms)

```

Figure 22.: Transitivity and totality together ($\exists\forall$ rule)

```

161 theorem T4:
162   assumes transitivity
163   shows Quasitransit
164   (*sledgehammer*)
165   by (metis assfactor_def assms)

```

Figure 23.: Proving quasi-transitivity

validity of the axiom, but not a necessary one. For instance, to disprove the implication “CM \Rightarrow m-smoothness” under the max rule (Fig. 12), *Nitpick* exhibits a model in which CM holds and m-smoothness is falsified. This model is shown in Fig. 24. 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

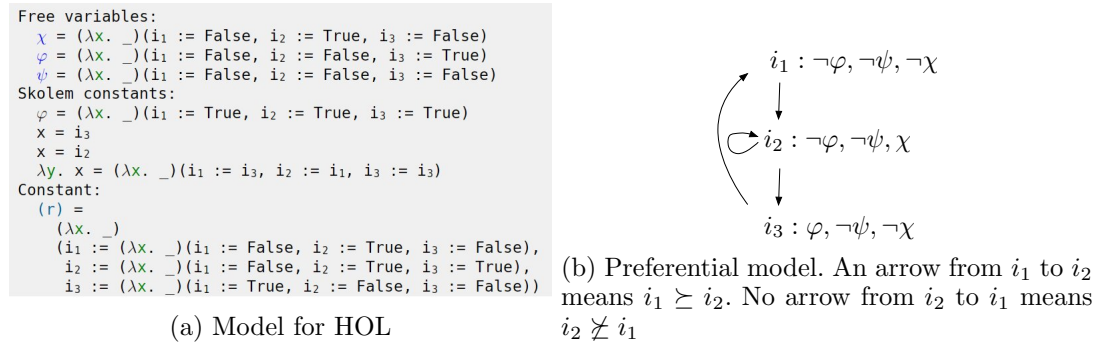


Figure 24.: A non-smooth model validating CM (max)

axiom. Also under both the max rule and the opt rule neither reflexivity nor totality correspond to an axiom. Finally, under the $\exists\forall$ rule the limit assumption has no impact. 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. Case study: Parfit’s repugnant conclusion

In this section,⁹ we show how to employ the framework described in the previous sections for the computer-aided assessment of ethical arguments in philosophy. Our focus is on analyzing the repugnant conclusion as discussed by Parfit (1984). We provide a computer encoding of his argument for the repugnant conclusion to make it amenable to formal analysis and computer-assisted experiments. Through the use of Isabelle/HOL, we discuss the plausibility of a solution of the paradox, advocated by Temkin (1987) and others. It involves rejecting the assumption of transitivity of “better than”. To put the proposed solution to the test, a full-blooded logical characterization of “better than” is needed. This one is given by the framework described in the previous sections. Following the tradition in deontic and conditional logic (refer to, for example, Lewis (1973)), we make a distinction between “better than” as a relation on formulas and as a relation on possible worlds, with the latter being instrumental in defining the logic of the former.

This section is organized into three subsections. Subsection 5.1 describes Parfit’s argument for the repugnant conclusion. For readability’s sake, we focus on a simplified version of the paradox, called the mere addition paradox. Subsection 5.2 documents the experiments we have run. Subsection 5.3 summarizes our findings.

5.1. The paradox

The repugnant conclusion reads:

“For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living.” (Parfit, 1984, Ch. 6)

The target is “total utilitarianism”, according to which the best outcome is given by the total of well-being in it. 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. 25 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.

⁹The code snippets are taken from the companion theory files “mere_addition_opt.thy”, “mere_addition_max.thy” and “mere_addition_lewis.thy”.

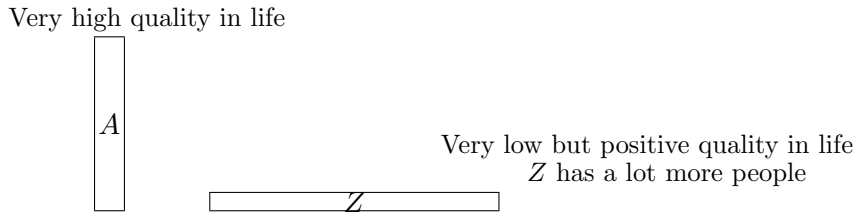


Figure 25.: Repugnant conclusion

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. 26. 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.

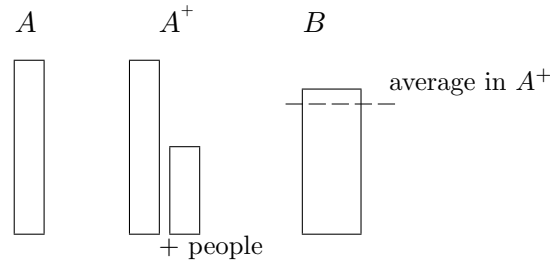


Figure 26.: Mere addition paradox

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^+ , ...) 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. 27 shows the encoding of P0-P2 in terms of obligation, an obligation statement being evaluated using the opt rule:

```

7 consts A::σ Aplus::σ B::σ
8
9 (*the mere addition scenario*)
10
11 (** With optimality **)
12
13
14 axiomatization where
15 (* A is strictly better than B*)
16 P0: "[¬⊙<¬A|AVB> ∧ ⊙<¬B|AVB>]" and
17 (* Aplus is at least as good as A*)
18 P1: "[¬⊙<¬Aplus|AVAplus>]" and
19 (* B is strictly better than Aplus*)
20 P2: "[¬⊙<¬B|AplusVB> ∧ ⊙<¬Aplus|AplusVB>]"

```

Figure 27.: Encoding of the mere addition scenario (optimality)

```

22 (* Sledgehammer finds P0-P2 inconsistent given
23 transitivity of the betterness relation in the models*)
24
25 theorem T0:
26   assumes transitivity
27   shows False
28   using P0 P1 P2 assms
29   sledgehammer
30   by (metis P0 P1 P2 assms)
31
32
33 (* Nitpick shows consistency in the absence of transitivity*)
34
35 theorem T1:
36   True
37   nitpick [satisfy, card i=3] (*model found*)
38   oops

```

Figure 28.: Sample queries on (P0)-(P2)

5.2. Computer-assisted experiments

Fig. 28 shows some sample queries run on the scenario under the opt rule. On l. 26, the assumption of the transitivity of the betterness relation (on possible worlds) is introduced. *Sledgehammer* shows the inconsistency of (P0)-(P2). On l. 35, the assumption of transitivity is dropped. *Nitpick* confirms the satisfiability of (P0)-(P2). The model generated by *Nitpick* is shown in Fig. 29.


```

Nitpicking formula...
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 := True, i3 := True)
Aplus = (λx. _)(i1 := False, i2 := True, i3 := False)
B = (λx. _)(i1 := True, i2 := False, i3 := False)

```

Figure 29.: A non-transitive model satisfying (P0)-(P2)

```

48 (* Nitpick shows consistency if transitivity is weakened into acyclicity or quasi-transitivity*)
49
50 theorem T3:
51   assumes loopfree
52   shows True
53   nitpick [show_all,satisfy,card=3] (* model found for card i=3 *)
54   oops
55
56 theorem T4:
57   assumes Quasitransit
58   shows True
59   nitpick [show_all,satisfy,card=4] (* model found for card i=4 *)
60   oops

```

Figure 30.: A-cyclicity and quasi-transitivity

```

40 (* Sledgehammer confirms inconsistency in the presence of the interval order condition*)
41
42 theorem T2:
43   assumes reflexivity Ferrers
44   shows False
45   sledgehammer
46   by (metis P0 P1 P2 assms(2))

```

Figure 31.: Interval order

Nitpick can also 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. 30 and 31.

We run the same queries under the max rule and the $\exists\forall$ rule. The findings are summarized in Table 4. The left-most column shows the constraint put on the betterness relation. The other columns show what happens when varying the truth conditions for the conditional obligation operator. The symbol \checkmark indicates that the sentences formalizing the scenario have been confirmed to be consistent, and the symbol \times indicates they have been confirmed to be inconsistent.

Property \ Truth conditions	Opt	Max	$\exists\forall$
None	✓	✓	✓
Transitivity + totality	✗	✗	✗
Transitivity	✗	✗ (if model finite)	✗
Interval order	✗	✗	✗
Quasi-transitivity	✓	✗ (if model finite)	✓
Acyclicity	✓	✓	✓

Table 4.: Mere addition paradox (overview of findings)

We verified manually the counter-models found by *Nitpick*, and all appeared to be correct. One can see that changing the truth conditions for the conditional does not have any effect, except for transitivity and quasi-transitivity under the max rule. A few comments are in order.

The formulas involved in the scenario are labeled as PP0-PP2. First of all, *Sledgehammer* shows that, if \succeq is transitive and total, then PP0-PP2 are inconsistent. This is shown in Fig. 32. This is to be contrasted with the situation under the opt rule and the $\exists\forall$ rule. But this apparent asymmetry has an explanation. When Temkin refers to the betterness relation, he has mostly in mind the relation \geq (on formulas). He does not disentangle the relation \geq (on formulas) from the relation \succeq (on worlds), the latter being used to define the truth conditions of the former. Nor does he specify if “best” is to be understood in terms of maximality or optimality. The properties of \succeq and \geq may not coincide depending on the definition of “best”. Thus, under the opt rule, if \succeq is transitive, then \geq is transitive—see T16 in Fig. 16. Under the max rule, it is only if \succeq is *both* transitive and total that \geq is transitive—see T12 in Fig. 15.¹⁰

```

99 theorem T0':
100   assumes transitivity and totality
101   shows False
102   sledgehammer
103   by (metis PP0 PP1 PP2 assms(1) assms(2))

```

Figure 32.: Inconsistency under transitivity and totality

However, this does not explain everything. The above might suggest an alternative solution to the mere addition paradox. Perhaps one could just keep the transitivity of \succeq but reject the totality of \succeq , while concurrently defining “best” in terms of maximality. This would be in keeping with the conventional approach in rational choice theory: maximality is often deemed more suitable than optimality, because it keeps the possibility of incomparability open—Sen (1997). But what happens with transitivity or quasi-transitivity of \succeq alone (third and fifth row, starting from the top) suggests that the solution lies elsewhere. *Sledgehammer* shows that, given (quasi-)transitivity, the formulas PP0-PP2 are inconsistent, assuming a finite model of cardinality (up to) seven (if we provide the exact dependencies). This is shown in Fig. 33.¹¹

Thus, under the max rule, (quasi-)transitivity makes PP0-PP2 inconsistent if the set of possible worlds is assumed to be finite—an assumption that might appear overly

¹⁰A pen and paper verification may be found in (Parent, 2021, Obs. 2.11 and 2.12).

¹¹For higher cardinalities *Sledgehammer* and *Nitpick* return a timeout. Timeouts can be explicitly specified by a parameter, like [timeout=100]; the default is 60s.

```

59 theorem T4:
60   assumes
61     transitivity and
62     OnlyOnes: "∀y. y=i1 ∨ y=i2 ∨ y=i3 ∨ y=i4 ∨ y= i5 ∨ y= i6 ∨ y= i7"
63   shows False
64   sledgehammer(PP0 PP1 PP2 assms assfactor_def)
65   oops
66
67 theorem T5:
68   assumes
69     Quasitransit and
70     OnlyOnes: "∀y. y=i1 ∨ y=i2 ∨ y=i3 ∨ y=i4 ∨ y= i5 ∨ y= i6 ∨ y=i7"
71   shows False
72   sledgehammer(PP0 PP1 PP2 assms assfactor_def)
73   oops

```

Figure 33.: Inconsistency under (quasi-)transitivity and finiteness

limiting, if not arbitrary. This observation calls into question the idea that transitivity is the sole cause of the paradox.

The above fact has remained unnoticed until now. We have not been able to establish it in full generality (i.e., regardless of the model's fixed cardinality) without resorting to manual calculations. This is Proposition 5.1 below. It says that, in the presence of (quasi-)transitivity, a necessary condition for PP0-PP2 to be simultaneously satisfiable, is that the model contains an infinite increasing \succ -chain of $\neg A \wedge \neg A^+ \wedge B$ -worlds (all distinct).

Proposition 5.1. *Suppose \succeq is quasi-transitive (resp. transitive). Assume the following formulas are satisfied in a world in a model $M = (W, \succeq, v)$:*

$$P(A/A \vee B) \tag{2}$$

$$P(A^+/A \vee A^+) \tag{3}$$

$$\bigcirc (\neg A^+/A^+ \vee B) \tag{4}$$

$$\bigcirc (\neg B/A \vee B) \tag{5}$$

$$P(B/A^+ \vee B) \tag{6}$$

Then W contains an infinite increasing \succ -chain of $\neg A \wedge \neg A^+ \wedge B$ -worlds (all distinct).

Proof. I focus on the case where \succeq is quasi-transitive. Recall that by definition \succ is irreflexive and asymmetric, and that quasi-transitivity entails acyclicity (see Fig. 1). Assume that formulas (3)-(4)-(5) are satisfied; the argument primarily revolves around these. Equations (2) and (6) can also be assumed true without leading to a contradiction.

By (3), there is some a_1 such that $a_1 \in \max(A \vee A^+)$ and $a_1 \models A^+$. Hence, $a_1 \models A^+ \vee B$. By (4), $a_1 \notin \max(A^+ \vee B)$, so there is some $a_2 \succ a_1$ s.t. $a_2 \models A^+ \vee B$. By irreflexivity, a_1 and a_2 are distinct. Clearly, $a_2 \models \neg A \wedge \neg A^+$. So $a_2 \models B$, and hence by (5), $a_2 \notin \max(A \vee B)$. It follows that there is some $a_3 \succ a_2$ s.t. $a_3 \models A \vee B$. By irreflexivity and asymmetry, a_3 is other than a_2 and a_1 . By quasi-transitivity, $a_3 \succ a_1$. Since $a_1 \in \max(A \vee A^+)$, $a_3 \models \neg A \wedge \neg A^+$. Hence $a_3 \models B$, and hence by (5) again, $a_3 \notin \max(A \vee B)$, and so there is some a_4 s.t. $a_4 \models A \vee B$ and $a_4 \succ a_3$. By acyclicity, a_4 is other than a_1, a_2, a_3 and a_4 . By quasi-transitivity, $a_4 \succ a_1$, and so as before $a_4 \models \neg A \wedge \neg A^+$. Reiterating the above argument indefinitely, one gets an infinite

increasing \succ -chain of $\neg A \wedge \neg A^+ \wedge B$ -worlds, starting with a_2 .

The argument goes through if \succeq is required to be transitive. This is because transitivity implies quasi-transitivity (see Fig. 1). \square

Remark 1. Note that Prop. 5.1 does not apply to the interval order condition (even if this one implies quasi-transitivity as well). This is because (6) cannot simultaneously hold in a model where the interval order condition is satisfied. This can easily be verified. By (6), there is some $b_1 \in \max(A^+ \vee B)$ with $b_1 \models B$. By (5), $b_1 \notin \max(A \vee B)$. So there is some $b_2 \succ b_1$ with $b_2 \models A \vee B$. Clearly, $b_2 \models A \wedge \neg B \wedge \neg A^+$. We have $a_1 \not\geq a_2$ and $b_1 \not\geq b_2$. By Ferrers, $a_1 \not\geq b_2$ or $b_1 \not\geq a_2$. By totality, $b_2 \succ a_1$ or $a_2 \succ b_1$. The first contradicts the fact that $a_1 \in \max(A \vee A^+)$, while the second contradicts the fact that $b_1 \in \max(A^+ \vee B)$. This is shown in Fig. 34, where the two cases are indicated by a “or”.

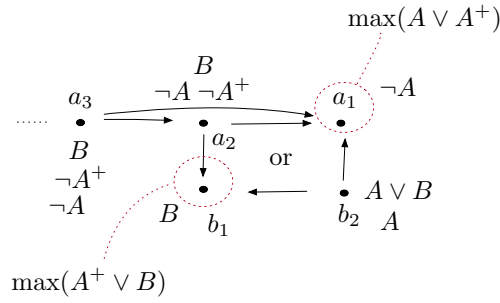


Figure 34.: Adding (6)

The following spin-off result is new to the literature. We recall that the finite model property (f.m.p.) is said to hold w.r.t. a given class C of models, if any formula φ that is satisfiable in class C is satisfiable in a finite model in C .

Corollary 5.2 (f.m.p.). *Under the max rule, the finite model property fails w.r.t. the following classes of models whose relation \succeq meets the property as indicated:*

- \succeq is quasi-transitive
- \succeq is transitive
- \succeq is an interval order

Proof. The second and third claims follow from the first, because quasi-transitivity follows from transitivity, and also from the interval order condition (see Fig. 1). To prove the first claim, set $\varphi := (3) \wedge (4) \wedge (5)$, and use Prop. 5.1 above. (The interval order condition will be met in the described model as long as each world is assumed to be at least as good as itself.) \square

In HOL one can define the axiom of infinity (for type i) by the second-order formula:

$$\text{infinity} \equiv \exists M. (\exists z :: i. \neg(Mz) \wedge (\exists G. (\forall y :: i. (\exists x. (Mx) \wedge (Gx = y))))))$$

The *definiens* says that there is a surjective mapping G from domain i to a proper subset M of domain i . Testing whether infinity holds, *Nitpick* gives us a counter-model to infinity that is a model of PP0-PP2. If the same query is run under the assumption of (quasi-)transitivity, we do not get any (finite) counter-model reported anymore.

```

78 abbreviation "infinity ≡ ∃M. (∃z::i. ¬(M z) ∧ (∃G. (∀y::i. (∃x. (M x) ∧ (G x) = y))))"
79
80 lemma "infinity" nitpick[show_all] oops (* countermodel found *)
81
82 (* Now we study infinity under the assumption of (quasi-)transitivity: we do
83 not get any finite countermodels reported anymore *)
84
85 lemma
86   assumes transitivity
87   shows infinity
88   nitpick[show_all] oops (* no countermodel found anymore; nitpicks runs out of time *)
89   sledgehammer (* but the provers are still too weak to prove it automatically *)
90
91 lemma
92   assumes Quasitransit
93   shows infinity
94   nitpick[show_all] (* no countermodel found anymore; nitpicks runs out of time *)
95   sledgehammer [max_proofs=1, isar_proofs=false] (* but the provers are still too
96   weak to prove it automatically *)
97   oops

```

Figure 35.: Proving infinity

However the provers are still not strong enough to prove infinity. This is shown in Fig. 35.

5.3. Summary

Distinguishing between “better than” as a relation between formulas and as a relation on possible worlds, our formalisation offers two new insights on the scenario. Below, it is understood that the structural properties are those of the second relation.

- One can choose not to take a stand on the truth conditions for the conditional, but weaken transitivity rather than reject it wholesale. However, not all potential weakenings of transitivity prove effective: quasi-transitivity and acyclicity do the job, but not the interval order condition. This is independent of the choice of the evaluation rule for the conditional.
- One could adopt the max rule, keep transitivity, and allow for the possibility that there are infinite sequences of better and better worlds. Ultimately, this solution is questionable, for the reason explained in Sect. 4.4: the max rule faces a deontic explosion problem, if infinite chains are allowed. Nevertheless, the availability of this option is worth a mention.

6. Conclusion

Utilising the LogiKEy methodology and framework we have developed mechanisations 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 formalisation and mechanisation 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 suitably. Future work includes the handling of the full equivalence between properties and formulas, the formalisation of (and comparison with) other solutions to the repugnant conclusion, and the analysis of other variant paradoxes discussed in the literature.

Funding

Dr. X. Parent was funded in whole, or in part, by the Austrian Science Fund (FWF) [M3240 N, ANCoR project (doi: 10.55776/I2982), and 10.55776/I6372, LoDEX project]. For the purpose of open access, the author has applied a CC BY public copy-right licence to any Author Accepted Manuscript version arising from this submission.

References

- Andrews, P. (2002). *An Introduction to Mathematical Logic and Type Theory*. Springer.
- Åqvist, L. (1987). *An Introduction to Deontic logic and the Theory of Normative Systems*. Bibliopolis, Naples.
- Åqvist, L. (2002). Deontic logic. In Gabbay, D. and Guenther, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, Dordrecht, Holland, 2nd edition. Originally published in (Gabbay and Guenther, 1984, pp. 605–714).
- 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*, 33(6):1243–1269.
- Benzmüller, C. and Woltzenlogel Paleo, B. (2016). The inconsistency in Gödel’s ontological argument: A success story for AI in metaphysics. In Kambhampati, S., editor, *IJCAI 2016*, volume 1-3, pages 936–942. AAAI Press.
- 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.
- Ciabattoni, A., Olivetti, N., and Parent, X. (2022). Dyadic obligations: Proofs and countermodels via hypersequents. In Aydoğan, R., Criado, N., Lang, J., Sánchez-Anguix, V.,

- and Serramia, M., editors, *PRIMA 2022: Principles and Practice of Multi-Agent Systems - 24th International Conference, Valencia, Spain, November 16-18, 2022, Proceedings*, volume 13753 of *Lecture Notes in Computer Science*, pages 54–71. Springer.
- Gabbay, D. and Guenther, F., editors (1984). *Handbook of Philosophical Logic*, volume II. Reidel, Dordrecht, Holland, 1st edition.
- Giordano, L., Gliozzi, V., and Pozzato, G. L. (2007). KLMLean 2.0: A theorem prover for KLM logics of nonmonotonic reasoning. In Olivetti, N., editor, *Automated Reasoning with Analytic Tableaux and Related Methods*, pages 238–244, Berlin, Heidelberg. Springer.
- Goble, L. (2004). A proposal for dealing with deontic dilemmas. In Lomuscio, A. and Nute, D., editors, *Deontic Logic in Computer Science*, pages 74–113, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Goble, L. (2019). Axioms for Hansson’s dyadic deontic logics. *Filosofiska Notiser*, 6(1):13–61.
- Hansson, B. (1969). An analysis of some deontic logics. *Noûs*, 3(4):373–398.
- Hughes, G. E. and Cresswell, M. J. (1968). *A Companion to Modal Logic*. Methuen, London.
- Kraus, S., Lehmann, D., and Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207.
- Lewis, D. (1973). *Counterfactuals*. Blackwell, Oxford.
- Luce, R. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191.
- Makinson, D. (1993). Five faces of minimality. *Studia Logica*, 52(3):339–379.
- Nortmann, U. (1986). Deontische logik: Die variante der lokalen äquivalenz. *Erkenntnis*, 6(25):275–318.
- 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. (2024). On some weakened forms of transitivity in the logic of conditional obligation. *Journal of Philosophical Logic*, 53(3):721–760.
- Parent, X. and Benzmüller, C. (2024). Conditional normative reasoning as a fragment of HOL (Isabelle/HOL dataset). *Archive of Formal Proofs*. Formal proof development, <https://isa-afp.org/entries/CondNormReasHOL.html>.
- 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.
- Sahlqvist, H. (1975). Completeness and correspondence in the first and second order semantics for modal logic. In Kanger, S., editor, *Proceedings of the Third Scandinavian Logic Symposium*, volume 82 of *Studies in Logic and the Foundations of Mathematics*, pages 110–143. Elsevier.
- Seegerberg, K. (1971). Some logics of commitment and obligation. In Hilpinen, R., editor, *Deontic Logic: Introductory and Systematic Readings*, pages 148–158, Dordrecht. Springer Netherlands.
- Sen, A. (1997). Maximization and the act of choice. *Econometrica*, 65(4):745–779.
- Shoham, Y. (1988). *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, MA, USA.
- Spohn, W. (1975). An analysis of Hansson’s dyadic deontic logic. *Journal of Philosophical Logic*, 4(2):237–252.
- Steen, A., Sutcliffe, G., Scholl, T., and Benzmüller, C. (2023). Solving modal logic problems by translation to higher-order logic. In Herzig, A., Luo, J., and Pardo, P., editors, *Logic and Argumentation*, pages 25–43, Cham. Springer Nature Switzerland.
- 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 Guentner, F., editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Springer Netherlands, Dordrecht.