

Conditional normative reasoning as a fragment of HOL (Isabelle/HOL dataset)

Xavier Parent and Christoph Benzmüller

March 29, 2024

Abstract

We present a mechanisation of (preference-based) conditional normative reasoning. Our focus is on Åqvist’s system **E** for conditional obligation and its extensions. We present both a correspondence-theory-focused metalogical study and a use-case application to Parfit’s repugnant conclusion, focusing on the mere addition paradox. Our contribution is explained in detail in [2]. This document presents a corresponding (but slightly modified) Isabelle/HOL dataset.

Contents

1	Introduction	2
2	Shallow Embedding of Åqvist’s system E	2
2.1	System E	2
2.2	Properties	4
3	Meta-Logical Study	6
3.1	Correspondence - Max rule	6
3.2	Correspondence - Opt Rule	9
3.3	Correspondence - Lewis’ rule	11
4	The Mere Addition Paradox: Opt Rule	14
5	The Mere Addition Paradox: Lewis’ rule	15
6	The Mere Addition Paradox: Max Rule	17
7	Conclusion	19

1 Introduction

In this document we present the Isabelle/HOL dataset associated with [2], in which “*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.*” [2]

2 Shallow Embedding of Åqvist’s system **E**

This is Åqvist’s system **E** from the 2019 IfColog paper [1].

2.1 System **E**

```
theory DDLcube
  imports Main
```

```
begin
```

```
nitpick-params [user-axioms,show-all,format=2] — Settings for model finder
Nitpick
```

```
typedecl i — Possible worlds
```

```
type-synonym  $\sigma = (i \Rightarrow \text{bool})$ 
```

```
type-synonym  $\alpha = i \Rightarrow \sigma$  — Type of betterness relation between worlds
```

```
type-synonym  $\tau = \sigma \Rightarrow \sigma$ 
```

```
consts aw::i — Actual world
```

```
abbreviation etrue ::  $\sigma (\top)$  where  $\top \equiv \lambda w. \text{True}$ 
```

```
abbreviation efalse ::  $\sigma (\perp)$  where  $\perp \equiv \lambda w. \text{False}$ 
```

```
abbreviation enot ::  $\sigma \Rightarrow \sigma$  ( $\neg$ -[52]53) where  $\neg\varphi \equiv \lambda w. \neg\varphi(w)$ 
```

```
abbreviation eand ::  $\sigma \Rightarrow \sigma \Rightarrow \sigma$  (infixr  $\wedge$  51) where  $\varphi \wedge \psi \equiv \lambda w. \varphi(w) \wedge \psi(w)$ 
```

```
abbreviation eor ::  $\sigma \Rightarrow \sigma \Rightarrow \sigma$  (infixr  $\vee$  50) where  $\varphi \vee \psi \equiv \lambda w. \varphi(w) \vee \psi(w)$ 
```

```
abbreviation eimpf ::  $\sigma \Rightarrow \sigma \Rightarrow \sigma$  (infixr  $\rightarrow$  49) where  $\varphi \rightarrow \psi \equiv \lambda w. \varphi(w) \rightarrow \psi(w)$ 
```

```
abbreviation eimpb ::  $\sigma \Rightarrow \sigma \Rightarrow \sigma$  (infixr  $\leftarrow$  49) where  $\varphi \leftarrow \psi \equiv \lambda w. \psi(w) \rightarrow \varphi(w)$ 
```

```
abbreviation eequ ::  $\sigma \Rightarrow \sigma \Rightarrow \sigma$  (infixr  $\leftrightarrow$  48) where  $\varphi \leftrightarrow \psi \equiv \lambda w. \varphi(w) \leftrightarrow \psi(w)$ 
```

abbreviation $ebox :: \sigma \Rightarrow \sigma$ (\Box) **where** $\Box\varphi \equiv \lambda w. \forall v. \varphi(v)$
abbreviation $ddediamond :: \sigma \Rightarrow \sigma$ (\Diamond) **where** $\Diamond\varphi \equiv \lambda w. \exists v. \varphi(v)$

abbreviation $evalid :: \sigma \Rightarrow bool$ ($[-]$ [8]109) — Global validity
where $[p] \equiv \forall w. p\ w$
abbreviation $ecjactual :: \sigma \Rightarrow bool$ ($[-]_i$ [7]105) — Local validity in world aw
where $[p]_i \equiv p(aw)$

consts $r :: \alpha$ (**infixr** r 70) — Betterness relation

abbreviation $esubset :: \sigma \Rightarrow \sigma \Rightarrow bool$ (**infix** \subseteq 53)
where $\varphi \subseteq \psi \equiv \forall x. \varphi\ x \longrightarrow \psi\ x$

We introduce the opt and max rules. These express two candidate truth-conditions for conditional obligation and permission.

abbreviation $eopt :: \sigma \Rightarrow \sigma$ ($opt<->$) — opt rule
where $opt<\varphi> \equiv (\lambda v. ((\varphi)(v) \wedge (\forall x. ((\varphi)(x) \longrightarrow v\ \mathbf{r}\ x))))$
abbreviation $econdopt :: \sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\odot<->$)
where $\odot<\psi|\varphi> \equiv \lambda w. opt<\varphi> \subseteq \psi$
abbreviation $eperm :: \sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\mathcal{P}<->$)
where $\mathcal{P}<\psi|\varphi> \equiv \neg\odot<\neg\psi|\varphi>$ — permission is the dual of obligation

abbreviation $emax :: \sigma \Rightarrow \sigma$ ($max<->$) — max rule
where $max<\varphi> \equiv (\lambda v. ((\varphi)(v) \wedge (\forall x. ((\varphi)(x) \longrightarrow (x\ \mathbf{r}\ v \longrightarrow v\ \mathbf{r}\ x)))))$
abbreviation $econd :: \sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\circ<->$)
where $\circ<\psi|\varphi> \equiv \lambda w. max<\varphi> \subseteq \psi$
abbreviation $euncobl :: \sigma \Rightarrow \sigma$ ($\circ<->$)
where $\circ<\varphi> \equiv \circ<\varphi|\top>$
abbreviation $ddeperm :: \sigma \Rightarrow \sigma \Rightarrow \sigma$ ($P<->$)
where $P<\psi|\varphi> \equiv \neg\circ<\neg\psi|\varphi>$

A first consistency check is performed.

lemma *True*
nitpick [*expect=genuine,satisfy*] — model found
<proof>

We show that the max -rule and opt -rule do not coincide.

lemma $\odot<\psi|\varphi> \equiv \circ<\psi|\varphi>$
nitpick [*expect=genuine,card i=1*] — counterexample found
<proof>

David Lewis's truth conditions for the deontic modalities are introduced.

abbreviation $lewcond :: \sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\circ<->$)
where $\circ<\psi|\varphi> \equiv \lambda v. (\neg(\exists x. (\varphi)(x)) \vee (\exists x. ((\varphi)(x) \wedge (\psi)(x) \wedge (\forall y. ((y\ \mathbf{r}\ x) \longrightarrow (\varphi)(y) \longrightarrow (\psi)(y))))))$
abbreviation $lewperm :: \sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\int<->$)
where $\int<\psi|\varphi> \equiv \neg\circ<\neg\psi|\varphi>$

Kratzer's truth conditions for the deontic modalities are introduced.

abbreviation *kratcond* :: $\sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\ominus \langle - | - \rangle$)
where $\ominus \langle \psi | \varphi \rangle \equiv \lambda v. ((\forall x. ((\varphi)(x) \longrightarrow$
 $(\exists y. ((\varphi)(y) \wedge (y \mathbf{r} x) \wedge ((\forall z. ((z \mathbf{r} y) \longrightarrow (\varphi)(z) \longrightarrow (\psi)(z))))))))))$
abbreviation *kratperm* :: $\sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\times \langle - | - \rangle$)
where $\times \langle \psi | \varphi \rangle \equiv \neg \ominus \langle \neg \psi | \varphi \rangle$

2.2 Properties

Extensions of **E** are obtained by putting suitable constraints on the betterness relation.

These are the standard properties of the betterness relation.

abbreviation *reflexivity* $\equiv (\forall x. x \mathbf{r} x)$
abbreviation *transitivity* $\equiv (\forall x y z. (x \mathbf{r} y \wedge y \mathbf{r} z) \longrightarrow x \mathbf{r} z)$
abbreviation *totality* $\equiv (\forall x y. (x \mathbf{r} y \vee y \mathbf{r} x))$

4 versions of Lewis's limit assumption can be distinguished.

abbreviation *mlimitedness* $\equiv (\forall \varphi. (\exists x. (\varphi)x) \longrightarrow (\exists x. \max \langle \varphi \rangle x))$

abbreviation *msmoothness* \equiv
 $(\forall \varphi x. ((\varphi)x \longrightarrow (\max \langle \varphi \rangle x \vee (\exists y. (y \mathbf{r} x \wedge \neg(x \mathbf{r} y) \wedge \max \langle \varphi \rangle y))))))$

abbreviation *olimitedness* $\equiv (\forall \varphi. (\exists x. (\varphi)x) \longrightarrow (\exists x. \text{opt} \langle \varphi \rangle x))$

abbreviation *osmoothness* \equiv
 $(\forall \varphi x. ((\varphi)x \longrightarrow (\text{opt} \langle \varphi \rangle x \vee (\exists y. (y \mathbf{r} x \wedge \neg(x \mathbf{r} y) \wedge \text{opt} \langle \varphi \rangle y))))))$

Weaker forms of transitivity can be defined. They require the notion of transitive closure.

definition *transitive* :: $\alpha \Rightarrow \text{bool}$
where *transitive Rel* $\equiv \forall x y z. \text{Rel } x y \wedge \text{Rel } y z \longrightarrow \text{Rel } x z$

definition *sub-rel* :: $\alpha \Rightarrow \alpha \Rightarrow \text{bool}$
where *sub-rel Rel1 Rel2* $\equiv \forall u v. \text{Rel1 } u v \longrightarrow \text{Rel2 } u v$

definition *assfactor* :: $\alpha \Rightarrow \alpha$
where *assfactor Rel* $\equiv \lambda u v. \text{Rel } u v \wedge \neg \text{Rel } v u$

In HOL the transitive closure of a relation can be defined in a single line - Here we apply the construction to the betterness relation and its strict variant.

definition *tcr*
where *tcr* $\equiv \lambda x y. \forall Q. \text{transitive } Q \longrightarrow (\text{sub-rel } r \ Q \longrightarrow Q \ x \ y)$

definition *tcr-strict*
where *tcr-strict* $\equiv \lambda x y. \forall Q. \text{transitive } Q$
 $\longrightarrow (\text{sub-rel } (\lambda u v. u \mathbf{r} v \wedge \neg v \mathbf{r} u) \ Q \longrightarrow Q \ x \ y)$

Quasi-transitivity requires the strict betterness relation is transitive.

abbreviation *Quasitransit*

where $Quasitransit \equiv \forall x y z. (assfactor\ r\ x\ y \wedge$
 $assfactor\ r\ y\ z) \longrightarrow assfactor\ r\ x\ z$

Suzumura consistency requires that cycles with at least one non-strict betterness link are ruled out.

abbreviation *Suzumura*

where $Suzumura \equiv \forall x y. tcr\ x\ y \longrightarrow (y\ \mathbf{r}\ x \longrightarrow x\ \mathbf{r}\ y)$

theorem *T1*: $Suzumura \equiv \forall x y. tcr\ x\ y \longrightarrow \neg (y\ \mathbf{r}\ x \wedge \neg (x\ \mathbf{r}\ y))$ *<proof>*

Acyclicity requires that cycles where all the links are strict are ruled out.

abbreviation *loopfree*

where $loopfree \equiv \forall x y. tcr\text{-}strict\ x\ y \longrightarrow (y\ \mathbf{r}\ x \longrightarrow x\ \mathbf{r}\ y)$

Interval order is the combination of reflexivity and Ferrers.

abbreviation *Ferrers*

where $Ferrers \equiv (\forall x y z u. (x\ \mathbf{r}\ u \wedge y\ \mathbf{r}\ z) \longrightarrow (x\ \mathbf{r}\ z \vee y\ \mathbf{r}\ u))$

theorem *T2*:

assumes *Ferrers* **and** *reflexivity* — fact overlooked in the literature

shows *totality*

— sledgehammer

<proof>

We study the relationships between these candidate weakenings of transitivity.

theorem *T3*:

assumes *transitivity*

shows *Suzumura*

— sledgehammer

<proof>

theorem *T4*:

assumes *transitivity*

shows *Quasitransit*

— sledgehammer

<proof>

theorem *T5*:

assumes *Suzumura*

shows *loopfree*

— sledgehammer

<proof>

theorem *T6*:

assumes *Quasitransit*

shows *loopfree*
— sledgehammer
<proof>

theorem *T7*:
assumes *reflexivity and Ferrers*
shows *Quasitransit*
— sledgehammer
<proof>

3 Meta-Logical Study

3.1 Correspondence - Max rule

The inference rules of **E** preserve validity in all models.

lemma *MP*: $[[\varphi]; [\varphi \rightarrow \psi]] \Longrightarrow [\psi]$
— sledgehammer
<proof>

lemma *NEC*: $[\varphi] \Longrightarrow [\Box\varphi]$
— sledgehammer
<proof>

\Box is an S5 modality

lemma *C-1-refl*: $[\Box\varphi \rightarrow \varphi]$
— sledgehammer
<proof>

lemma *C-1-trans*: $[\Box\varphi \rightarrow (\Box(\Box\varphi))]$
— sledgehammer
<proof>

lemma *C-1-sym*: $[\varphi \rightarrow (\Box(\Diamond\varphi))]$
— sledgehammer
<proof>

All the axioms of **E** hold - they do not correspond to a property of the betterness relation.

lemma *Abs*: $[\Box\langle\psi|\varphi\rangle \rightarrow \Box\Box\langle\psi|\varphi\rangle]$
— sledgehammer
<proof>

lemma *Nec*: $[\Box\psi \rightarrow \Box\langle\psi|\varphi\rangle]$
— sledgehammer
<proof>

lemma *Ext*: $[\Box(\varphi_1 \leftrightarrow \varphi_2) \rightarrow (\Box\langle\psi|\varphi_1\rangle \leftrightarrow \Box\langle\psi|\varphi_2\rangle)]$
— sledgehammer

<proof>

lemma *Id*: $[\bigcirc\langle\varphi|\varphi\rangle]$
— sledgehammer
<proof>

lemma *Sh*: $[\bigcirc\langle\psi|\varphi_1\wedge\varphi_2\rangle \rightarrow \bigcirc\langle(\varphi_2\rightarrow\psi)|\varphi_1\rangle]$
— sledgehammer
<proof>

lemma *COK*: $[\bigcirc\langle(\psi_1\rightarrow\psi_2)|\varphi\rangle \rightarrow (\bigcirc\langle\psi_1|\varphi\rangle \rightarrow \bigcirc\langle\psi_2|\varphi\rangle)]$
— sledgehammer
<proof>

The axioms of the stronger systems do not hold in general.

lemma $[\diamond\varphi \rightarrow (\bigcirc\langle\psi|\varphi\rangle \rightarrow P\langle\psi|\varphi\rangle)]$
nitpick [*expect=genuine, card i=3*] — counterexample found
<proof>

lemma $[(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\chi|\varphi\rangle) \rightarrow \bigcirc\langle\chi|\varphi\wedge\psi\rangle]$
nitpick [*expect=genuine, card i=3*] — counterexample found
<proof>

lemma $[\bigcirc\langle\chi|(\varphi\vee\psi)\rangle \rightarrow ((\bigcirc\langle\chi|\varphi\rangle) \vee (\bigcirc\langle\chi|\psi\rangle))]$
nitpick [*expect=genuine, card i=3*] — counterexample found
<proof>

Now we identify a number of correspondences under the max rule. Only the direction property \Rightarrow axiom is verified.

Max-limitedness corresponds to D^* , the distinctive axiom of **F**. The first implies the second, but not the other around.

theorem *T8*:
assumes *mlimitedness*
shows D^* : $[\diamond\varphi \rightarrow \bigcirc\langle\psi|\varphi\rangle \rightarrow P\langle\psi|\varphi\rangle]$
— sledgehammer
<proof>

lemma
assumes D^* : $[\diamond\varphi \rightarrow \neg(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\neg\psi|\varphi\rangle)]$
shows *mlimitedness*
nitpick [*expect=genuine, card i=3*] — counterexample found
<proof>

Smoothness implies cautious monotony, the distinctive axiom of **F**+(CM), but not the other way around.

theorem *T9*:
assumes *msmoothness*
shows *CM*: $[(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\chi|\varphi\rangle) \rightarrow \bigcirc\langle\chi|\varphi\wedge\psi\rangle]$

— sledgehammer
<proof>

lemma

assumes *CM*: $[(\bigcirc\langle\psi|\varphi\rangle \wedge \bigcirc\langle\chi|\varphi\rangle) \rightarrow \bigcirc\langle\chi|\varphi\wedge\psi\rangle]$

shows *msmoothness*

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

Interval order corresponds to disjunctive rationality, the distinctive axiom of $\mathbf{F}+(\text{DR})$, but not the other way around.

lemma

assumes *reflexivity*

shows *DR*: $[\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)]$

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

theorem *T10*:

assumes *reflexivity and Ferrers*

shows *DR*: $[\bigcirc\langle\chi|(\varphi\vee\psi)\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)]$

— sledgehammer

<proof>

lemma

assumes *DR*: $[\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)]$

shows *reflexivity*

nitpick [*expect=genuine, card i=1*] — counterexample found

<proof>

lemma

assumes *DR*: $[\bigcirc\langle\chi|\varphi\vee\psi\rangle \rightarrow (\bigcirc\langle\chi|\varphi\rangle \vee \bigcirc\langle\chi|\psi\rangle)]$

shows *Ferrers*

nitpick [*expect=genuine, card i=2*] — counterexample found

<proof>

Transitivity and totality jointly correspond to the Spohn axiom (Sp), the distinctive axiom of system \mathbf{G} , but not vice-versa. They also jointly correspond to a principle of transitivity for the betterness relation on formulas, but the converse fails.

lemma

assumes *transitivity*

shows *Sp*: $[(P\langle\psi|\varphi\rangle \wedge \bigcirc\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \bigcirc\langle\chi|(\varphi\wedge\psi)\rangle]$

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

lemma

assumes *totality*

shows *Sp*: $[(P\langle\psi|\varphi\rangle \wedge \bigcirc\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \bigcirc\langle\chi|(\varphi\wedge\psi)\rangle]$

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

theorem T11:

assumes *transitivity and totality*

shows *Sp*: $[(P\langle\psi|\varphi\rangle \wedge O\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow O\langle\chi|(\varphi\wedge\psi)\rangle]$

— sledgehammer

<proof>

theorem T12:

assumes *transitivity and totality*

shows *transit*: $[(P\langle\varphi|\varphi\vee\psi\rangle \wedge P\langle\psi|\psi\vee\chi\rangle) \rightarrow P\langle\varphi|(\varphi\vee\chi)\rangle]$

— sledgehammer

<proof>

lemma

assumes *Sp*: $[(P\langle\psi|\varphi\rangle \wedge O\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow O\langle\chi|(\varphi\wedge\psi)\rangle]$

shows *totality*

nitpick [*expect=genuine, card i=1*] — counterexample found

<proof>

lemma

assumes *Sp*: $[(P\langle\psi|\varphi\rangle \wedge O\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow O\langle\chi|(\varphi\wedge\psi)\rangle]$

shows *transitivity*

nitpick [*expect=genuine, card i=2*] — counterexample found

<proof>

3.2 Correspondence - Opt Rule

Opt-limitedness corresponds to D, but not vice-versa.

theorem T13:

assumes *olimitedness*

shows *D*: $[\Diamond\varphi \rightarrow \odot\langle\psi|\varphi\rangle \rightarrow \mathcal{P}\langle\psi|\varphi\rangle]$

— sledgehammer

<proof>

lemma

assumes *D*: $[\Diamond\varphi \rightarrow \odot\langle\psi|\varphi\rangle \rightarrow \mathcal{P}\langle\psi|\varphi\rangle]$

shows *olimitedness*

nitpick [*expect=genuine, card i=1*] — counterexample found

<proof>

Smoothness implies cautious monotony, but not vice-versa.

theorem T14:

assumes *smoothness*

shows *CM'*: $[(\odot\langle\psi|\varphi\rangle \wedge \odot\langle\chi|\varphi\rangle) \rightarrow \odot\langle\chi|\varphi\wedge\psi\rangle]$

— sledgehammer

<proof>

lemma

assumes *CM*: $[(\odot\langle\psi|\varphi\rangle \wedge \odot\langle\chi|\varphi\rangle) \rightarrow \odot\langle\chi|\varphi\wedge\psi\rangle]$
shows *smoothness*
nitpick [*expect=genuine, card i=1*] — counterexample found
 $\langle proof \rangle$

Transitivity (on worlds) implies Sp and transitivity (on formulas), but not vice-versa.

theorem *T15*:

assumes *transitivity*
shows *Sp'*: $[(\mathcal{P}\langle\psi|\varphi\rangle \wedge \odot\langle\psi\rightarrow\chi|\varphi\rangle) \rightarrow \odot\langle\chi|(\varphi\wedge\psi)\rangle]$
— sledgehammer
 $\langle proof \rangle$

theorem *T16*:

assumes *transitivity*
shows *Trans'*: $[(\mathcal{P}\langle\varphi|\varphi\vee\psi\rangle \wedge \mathcal{P}\langle\psi|\psi\vee\xi\rangle) \rightarrow \mathcal{P}\langle\varphi|\varphi\vee\xi\rangle]$
— sledgehammer
 $\langle proof \rangle$

lemma

assumes *Sp*: $[(\mathcal{P}\langle\psi|\varphi\rangle \wedge \odot\langle\psi\rightarrow\chi|\varphi\rangle) \rightarrow \odot\langle\chi|(\varphi\wedge\psi)\rangle]$
assumes *Trans*: $[(\mathcal{P}\langle\varphi|\varphi\vee\psi\rangle \wedge \mathcal{P}\langle\psi|\psi\vee\xi\rangle) \rightarrow \mathcal{P}\langle\varphi|\varphi\vee\xi\rangle]$
shows *transitivity*
nitpick [*expect=genuine, card i=2*] — counterexample found
 $\langle proof \rangle$

Interval order implies disjunctive rationality, but not vice-versa.

lemma

assumes *reflexivity*
shows *DR'*: $[\odot\langle\chi|\varphi\vee\psi\rangle \rightarrow (\odot\langle\chi|\varphi\rangle \vee \odot\langle\chi|\psi\rangle)]$
nitpick [*expect=genuine, card i=3*] — counterexample found
 $\langle proof \rangle$

theorem *T17*:

assumes *reflexivity and Ferrers*
shows *DR'*: $[\odot\langle\chi|\varphi\vee\psi\rangle \rightarrow (\odot\langle\chi|\varphi\rangle \vee \odot\langle\chi|\psi\rangle)]$
— sledgehammer
 $\langle proof \rangle$

lemma

assumes *DR*: $[\odot\langle\chi|\varphi\vee\psi\rangle \rightarrow (\odot\langle\chi|\varphi\rangle \vee \odot\langle\chi|\psi\rangle)]$
shows *reflexivity*
nitpick [*expect=genuine, card i=1*] — counterexample found
 $\langle proof \rangle$

lemma

assumes *DR*: $[\odot\langle\chi|\varphi\vee\psi\rangle \rightarrow (\odot\langle\chi|\varphi\rangle \vee \odot\langle\chi|\psi\rangle)]$
shows *Ferrers*
nitpick [*expect=genuine, card i=2*] — counterexample found

<proof>

3.3 Correspondence - Lewis' rule

We have deontic explosion under the max rule.

theorem *DEX*: $[(\Diamond\varphi \wedge \circ\langle\psi|\varphi\rangle \wedge \circ\langle\neg\psi|\varphi\rangle) \rightarrow \circ\langle\chi|\varphi\rangle]$

— sledgehammer

<proof>

But no deontic explosion under Lewis' rule.

lemma *DEX*: $[(\Diamond\varphi \wedge \circ\langle\psi|\varphi\rangle \wedge \circ\langle\neg\psi|\varphi\rangle) \rightarrow \circ\langle\chi|\varphi\rangle]$

nitpick [*expect=genuine, card i=2*] — counterexample found

<proof>

The three rules are equivalent when the betterness relation meets all the standard properties.

theorem *T18*:

assumes *mlimitedness and transitivity and totality*

shows $[\circ\langle\psi|\varphi\rangle \leftrightarrow \circ\langle\psi|\varphi\rangle]$

— sledgehammer

<proof>

theorem *T19*:

assumes *mlimitedness and transitivity and totality*

shows $[\circ\langle\psi|\varphi\rangle \leftrightarrow \circ\langle\psi|\varphi\rangle]$

— sledgehammer

<proof>

These are the axioms of **E** that do not call for a property.

theorem *Abs'*: $[\circ\langle\psi|\varphi\rangle \rightarrow \Box\circ\langle\psi|\varphi\rangle]$

— sledgehammer

<proof>

theorem *Nec'*: $[\Box\psi \rightarrow \circ\langle\psi|\varphi\rangle]$

— sledgehammer

<proof>

theorem *Ext'*: $[\Box(\varphi_1 \leftrightarrow \varphi_2) \rightarrow (\circ\langle\psi|\varphi_1\rangle \leftrightarrow \circ\langle\psi|\varphi_2\rangle)]$

— sledgehammer

<proof>

theorem *Id'*: $[\circ\langle\varphi|\varphi\rangle]$

— sledgehammer

<proof>

theorem *Sh'*: $[\circ\langle\psi|\varphi_1 \wedge \varphi_2\rangle \rightarrow \circ\langle(\varphi_2 \rightarrow \psi)|\varphi_1\rangle]$

— sledgehammer

<proof>

One axiom of **E**, and the distinctive axioms of its extensions are invalidated in the absence of a property of the betterness relation.

lemma D: $[\Diamond\varphi \rightarrow (\circ\langle\psi|\varphi\rangle \rightarrow \int\langle\psi|\varphi\rangle)]$
nitpick [*expect=genuine, card i=2*] — counterexample found
 $\langle proof \rangle$

lemma Sp: $[(\int\langle\psi|\varphi\rangle \wedge \circ\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \circ\langle\chi|(\varphi\wedge\psi)\rangle]$
nitpick [*expect=genuine, card i=3*] — counterexample found
 $\langle proof \rangle$

lemma COK: $[\circ\langle(\psi_1\rightarrow\psi_2)|\varphi\rangle \rightarrow (\circ\langle\psi_1|\varphi\rangle \rightarrow \circ\langle\psi_2|\varphi\rangle)]$
nitpick [*expect=genuine, card i=2*] — counterexample found
 $\langle proof \rangle$

lemma CM: $[(\circ\langle\psi|\varphi\rangle \wedge \circ\langle\chi|\varphi\rangle) \rightarrow \circ\langle\chi|\varphi\wedge\psi\rangle]$
nitpick [*expect=genuine, card i=2*] — counterexample found
 $\langle proof \rangle$

Totality implies the distinctive axiom of **F**, but not vice-versa.

theorem T20:
assumes *totality*
shows $[\Diamond\varphi \rightarrow (\circ\langle\psi|\varphi\rangle \rightarrow \int\langle\psi|\varphi\rangle)]$
— sledgehammer
 $\langle proof \rangle$

lemma
assumes $[\Diamond\varphi \rightarrow (\circ\langle\psi|\varphi\rangle \rightarrow \int\langle\psi|\varphi\rangle)]$
shows *totality*
nitpick [*expect=genuine, card i=3*] — counterexample found
 $\langle proof \rangle$

Transitivity implies the distinctive axioms of **G**, but not vice-versa.

theorem T21:
assumes *transitivity*
shows Sp'' : $[(\int\langle\psi|\varphi\rangle \wedge \circ\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \circ\langle\chi|(\varphi\wedge\psi)\rangle]$
— sledgehammer
 $\langle proof \rangle$

theorem T22:
assumes *transitivity*
shows Tr'' : $[(\int\langle\varphi|\varphi\vee\psi\rangle \wedge \int\langle\psi|\psi\vee\chi\rangle) \rightarrow \int\langle\varphi|\varphi\vee\chi\rangle]$
— sledgehammer
 $\langle proof \rangle$

lemma
assumes Sp'' : $[(\int\langle\psi|\varphi\rangle \wedge \circ\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \circ\langle\chi|(\varphi\wedge\psi)\rangle]$
shows *transitivity*
nitpick — counterexample found

<proof>

lemma

assumes Tr'' : $\lfloor (\int \langle \varphi | \varphi \vee \psi \rangle \wedge \int \langle \psi | \psi \vee \chi \rangle) \rightarrow \int \langle \varphi | \varphi \vee \chi \rangle \rfloor$

shows *transitivity*

nitpick — counterexample found

<proof>

lemma

assumes *transitivity*

shows COK : $\lfloor \circ \langle (\psi_1 \rightarrow \psi_2) | \varphi \rangle \rightarrow (\circ \langle \psi_1 | \varphi \rangle \rightarrow \circ \langle \psi_2 | \varphi \rangle) \rfloor$

nitpick [*expect=genuine, card i=2*] — counterexample found

<proof>

lemma

assumes *totality*

shows COK : $\lfloor \circ \langle (\psi_1 \rightarrow \psi_2) | \varphi \rangle \rightarrow (\circ \langle \psi_1 | \varphi \rangle \rightarrow \circ \langle \psi_2 | \varphi \rangle) \rfloor$

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

Transitivity and totality imply an axiom of **E** and the distinctive axiom of **F+CM**, but not vice-versa.

theorem *T23*:

assumes *transitivity and totality*

shows COK' : $\lfloor \circ \langle (\psi_1 \rightarrow \psi_2) | \varphi \rangle \rightarrow (\circ \langle \psi_1 | \varphi \rangle \rightarrow \circ \langle \psi_2 | \varphi \rangle) \rfloor$

— sledgehammer

<proof>

lemma

assumes COK' : $\lfloor \circ \langle (\psi_1 \rightarrow \psi_2) | \varphi \rangle \rightarrow (\circ \langle \psi_1 | \varphi \rangle \rightarrow \circ \langle \psi_2 | \varphi \rangle) \rfloor$

shows *transitivity and totality*

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

theorem *T24*:

assumes *transitivity and totality*

shows CM'' : $\lfloor (\circ \langle \psi | \varphi \rangle \wedge \circ \langle \chi | \varphi \rangle) \rightarrow \circ \langle \chi | \varphi \wedge \psi \rangle \rfloor$

— sledgehammer

<proof>

lemma

assumes CM'' : $\lfloor (\circ \langle \psi | \varphi \rangle \wedge \circ \langle \chi | \varphi \rangle) \rightarrow \circ \langle \chi | \varphi \wedge \psi \rangle \rfloor$

shows *transitivity and totality*

nitpick [*expect=genuine, card i=3*] — counterexample found

<proof>

Under the opt rule transitivity alone imply Sp and Trans, but not vice-versa.

theorem *T25*:

assumes *transitivity*

shows $[(\mathcal{P}\langle\psi|\varphi\rangle \wedge \odot\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \odot\langle\chi|(\varphi\wedge\psi)\rangle]$
— sledgehammer
 $\langle proof \rangle$

lemma

assumes *transitivity*
shows $[(\mathcal{P}\langle\varphi|\varphi\vee\psi\rangle \wedge \mathcal{P}\langle\xi|\psi\vee\xi\rangle) \rightarrow \mathcal{P}\langle\xi|\varphi\vee\xi\rangle]$
nitpick [*expect=genuine, card i=2*] — counterexample found
 $\langle proof \rangle$

lemma

assumes *Sp*: $[(\mathcal{P}\langle\psi|\varphi\rangle \wedge \odot\langle(\psi\rightarrow\chi)|\varphi\rangle) \rightarrow \odot\langle\chi|(\varphi\wedge\psi)\rangle]$
and *Trans*: $[(\mathcal{P}\langle\varphi|\varphi\vee\psi\rangle \wedge \mathcal{P}\langle\xi|\psi\vee\xi\rangle) \rightarrow \mathcal{P}\langle\xi|\varphi\vee\xi\rangle]$
shows *transitivity*
nitpick [*expect=genuine, card i=2*] — counterexample found
 $\langle proof \rangle$

end

4 The Mere Addition Paradox: Opt Rule

This section studies the mere addition paradox [3], when assuming the opt rule. The mere addition paradox is a smaller version of Parfit’s repugnant conclusion.

We assess the well-known solution advocated by e.g. Temkin [4] among others, which consists in abandoning the transitivity of the betterness relation.

theory *mere-addition-opt*
imports *DDLcube*

begin

consts *A::σ Aplus::σ B::σ*

Here is the formalization of the paradox.

axiomatization where

— A is strictly better than B

P0: $[(\neg\odot\langle\neg A|A\vee B\rangle \wedge \odot\langle\neg B|A\vee B\rangle)]$ **and**

— Aplus is at least as good as A

P1: $[\neg\odot\langle\neg Aplus|A\vee Aplus\rangle]$ **and**

— B is strictly better than Aplus

P2: $[(\neg\odot\langle\neg B|Aplus\vee B\rangle \wedge \odot\langle\neg Aplus|Aplus\vee B\rangle)]$

Sledgehammer finds P0-P2 inconsistent given transitivity of the betterness relation in the models:

theorem *T0*:

assumes *transitivity*

```
shows False
— sledgehammer
⟨proof⟩
```

Nitpick shows consistency in the absence of transitivity:

```
theorem T1:
  True
nitpick [satisfy, expect=genuine, card i=3] — model found
⟨proof⟩
```

Now we consider what happens when transitivity is weakened suitably rather than abandoned wholesale. We show that this less radical solution is also possible, but that not all candidate weakenings are effective.

Sledgehammer confirms inconsistency in the presence of the interval order condition:

```
theorem T2:
assumes reflexivity Ferrers
shows False
— sledgehammer
⟨proof⟩
```

Nitpick shows consistency if transitivity is weakened into acyclicity or quasi-transitivity:

```
theorem T3:
assumes loopfree
shows True
nitpick [satisfy, expect=genuine, card=3] — model found
⟨proof⟩
```

```
theorem T4:
assumes Quasitransit
shows True
nitpick [satisfy, expect=genuine, card=4] — model found
⟨proof⟩
```

end

5 The Mere Addition Paradox: Lewis' rule

We run the same queries as before, but using Lewis' rule. The outcome is pretty much the same. Thus, the choice between the opt rule and Lewis' rule does not make a difference.

```
theory mere-addition-lewis
imports DDLcube
```

begin

consts $a::\sigma$ $aplus::\sigma$ $b::\sigma$

axiomatization where

- A is strictly better than B
- $PPP0$: $[(\neg \circ < \neg a | a \vee b > \wedge \circ < \neg b | a \vee b >)]$ **and**
- Aplus is at least as good as A
- $PPP1$: $[(\neg \circ < \neg aplus | a \vee aplus >)]$ **and**
- B is strictly better than Aplus
- $PPP2$: $[(\neg \circ < \neg b | aplus \vee b > \wedge \circ < \neg aplus | aplus \vee b >)]$

Sledgehammer finds PPP0-PPP2 inconsistent given transitivity of the betterness relation in the models:

theorem $T0$:

- assumes** *transitivity*
- shows** *False*
- sledgehammer
- $\langle proof \rangle$

Nitpick shows consistency in the absence of transitivity:

lemma $T1$:

- True*
- nitpick** [*satisfy, expect=genuine, card i=3, show-all*] — model found
- $\langle proof \rangle$

Sledgehammer confirms inconsistency in the presence of the interval order condition:

theorem $T2$:

- assumes** *reflexivity Ferrers*
- shows** *False*
- sledgehammer
- $\langle proof \rangle$

Nitpick shows consistency if transitivity is weakened into acyclicity or quasi-transitivity:

theorem $T3$:

- assumes** *loopfree*
- shows** *True*
- nitpick** [*satisfy, expect=genuine, card=3*] — model found
- $\langle proof \rangle$

theorem $T4$:

- assumes** *Quasitransit*
- shows** *True*
- nitpick** [*satisfy, expect=genuine, card=4*] — model found
- $\langle proof \rangle$

end

6 The Mere Addition Paradox: Max Rule

There are surprising results with the max rule. Transitivity alone generates an inconsistency only when combined with totality. What is more, given transitivity (or quasi-transitivity) alone, the formulas turn out to be all satisfiable in an infinite model.

```

theory mere-addition-max
  imports DDLcube

begin

consts A:: $\sigma$  Aplus:: $\sigma$  B:: $\sigma$  i1::i i2::i i3::i i4::i i5::i i6::i i7::i i8::i

```

axiomatization where

```

— A is strictly better than B
PP0: [ $\neg \circ < \neg A | A \vee B > \wedge \circ < \neg B | A \vee B >$ ] and
— Aplus is at least as good as A
PP1: [ $\neg \circ < \neg Aplus | A \vee Aplus >$ ] and
— B is strictly better than Aplus
PP2: [ $\neg \circ < \neg B | Aplus \vee B > \wedge \circ < \neg Aplus | Aplus \vee B >$ ]

```

Nitpick finds no finite model when the betterness relation is assumed to be transitive:

```

theorem T0:
  assumes transitivity
  shows True
  nitpick [satisfy, expect=none] — no model found
  <proof>

```

Nitpick shows consistency in the absence of transitivity:

```

theorem T1:
  shows True
  nitpick [satisfy, expect=genuine, card i=3] — model found
  <proof>

```

Sledgehammer confirms inconsistency in the presence of the interval order condition:

```

theorem T2:
  assumes reflexivity and Ferrers
  shows False
  — sledgehammer
  <proof>

```

Nitpick shows consistency if transitivity is weakened into acyclicity:

```

theorem T3:
  assumes loopfree
  shows True

```

nitpick [*satisfy, expect=genuine, card=3*] — model found
(*proof*)

If transitivity or quasi-transitivity is assumed, Nitpick shows inconsistency assuming a finite model of cardinality (up to) seven (if we provide the exact dependencies)—for higher cardinalities it returns a time out (depending on the computer it may prove falsity also for cardinality eight, etc.:

theorem T4:

assumes

transitivity **and**

OnlyOnes: $\forall y. y=i1 \vee y=i2 \vee y=i3 \vee y=i4 \vee y=i5 \vee y=i6 \vee y=i7$

shows *False*

(*proof*)

theorem T5:

assumes

Quasitransit **and**

OnlyOnes: $\forall y. y=i1 \vee y=i2 \vee y=i3 \vee y=i4 \vee y=i5 \vee y=i6 \vee y=i7$

shows *False*

(*proof*)

Infinity is encoded as follows: there is a surjective mapping G from domain i to a proper subset M of domain i . Testing whether infinity holds in general Nitpick finds a countermodel:

abbreviation *infinity* $\equiv \exists M. (\exists z::i. \neg(M z) \wedge (\exists G. (\forall y::i. (\exists x. (M x) \wedge (G x) = y))))$

lemma *infinity nitpick*[*expect=genuine*] (*proof*)

Now we run the same query under the assumption of (quasi-)transitivity: we do not get any finite countermodel reported anymore:

lemma

assumes *transitivity*

shows *infinity*

— nitpick — no countermodel found anymore; nitpicks runs out of time

— sledgehammer — but the provers are still too weak to prove it automatically;

see [2] for a pen and paper proof

(*proof*)

lemma

assumes *Quasitransit*

shows *infinity*

— nitpick — no countermodel found anymore; nitpicks runs out of time

— sledgehammer — but the provers are still too weak to prove it automatically;

see [2] for a pen and paper proof

(*proof*)

Transitivity and totality together give inconsistency:

theorem *T0'*:
 assumes *transitivity and totality*
 shows *False*
 — sledgehammer
 ⟨*proof*⟩

end

7 Conclusion

In this document we presented the Isabelle/HOL dataset associated with [2]. We described our shallow semantic embedding of Åqvist’s dyadic deontic logic **E** and its extensions. We showcased two key uses of the framework: first, for meta-reasoning about the logic, particularly for verifying deontic correspondences similar to modal logic; second, for assessing ethical arguments, exemplified by encoding Parfit’s mere addition paradox, a smaller version of his so-called repugnant conclusion.

References

- [1] C. Benzmüller, A. Farjami, and X. Parent. Åqvist’s dyadic deontic logic E in HOL. *Journal of Applied Logics – IfCoLoG Journal of Logics and their Applications (Special Issue: Reasoning for Legal AI)*, 6(5):733–755, 2019.
- [2] X. Parent and C. Benzmüller. Conditional normative reasoning as a fragment of HOL. *Submitted for journal publication*, 2024. minor revisions, revision submitted, preprint: <https://arxiv.org/abs/2308.10686>.
- [3] D. Parfit. *Reasons and Persons*. Oxford University Press, 1984.
- [4] L. S. Temkin. Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16(2):138–187, 1987.