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

Xavier Parent and Christoph Benzmüller

May 26, 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 sligthly modified) Isabelle/HOL dataset.

Contents

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 Aqvist'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 $\sigma = (i \Rightarrow bool)$
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* :: σ $(\top)$ **where** $\top \equiv \lambda w.\ True$
abbreviation *efalse* :: σ $(\bot)$ **where** $\bot \equiv \lambda w.\ False$
abbreviation *enot* :: $\sigma \Rightarrow \sigma$ $(\neg\text{-}[52]53)$ **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** $\varphi\rightarrow\psi \equiv \lambda w.\ \varphi(w)\longrightarrow\psi(w)$

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

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

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

abbreviation *evalid* :: $\sigma \Rightarrow bool$ ($\lfloor \text{-} \rfloor [8] 109$) — Global validity
where $\lfloor p \rfloor \equiv \forall w.\ p\ w$
abbreviation *ecjactual* :: $\sigma \Rightarrow bool$ ($\lfloor \text{-} \rfloor_l [7] 105$) — Local validity in world aw
where $\lfloor p \rfloor_l \equiv p(aw)$

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

abbreviation *emax* :: $\sigma \Rightarrow \sigma$ ($max{<}\text{-}{>}$) — 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* :: $\sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\circ{<}\text{-}|\text{-}{>}$)
where $\circ{<}\psi|\varphi{>} \equiv\ \lambda w.\ max{<}\varphi{>} \subseteq \psi$
abbreviation *euncobl* :: $\sigma \Rightarrow \sigma$ ($\bigcirc{<}\text{-}{>}$)
where $\bigcirc{<}\varphi{>} \equiv \circ{<}\varphi|\top{>}$
abbreviation *ddeperm* :: $\sigma \Rightarrow \sigma \Rightarrow \sigma$ ($P{<}\text{-}|\text{-}{>}$)
where $P{<}\psi|\varphi{>} \equiv \neg\circ{<}\neg\psi|\varphi{>}$

A first consistency check is performed.

lemma *True*
nitpick [*expect=genuine,satisfy*] — model found
oops

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
oops

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

abbreviation *lewcond* :: $\sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\circ{<}\text{-}|\text{-}{>}$)
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* :: $\sigma \Rightarrow \sigma \Rightarrow \sigma$ ($\int{<}\text{-}|\text{-}{>}$)
where $\int{<}\psi|\varphi{>} \equiv \neg\circ{<}\neg\psi|\varphi{>}$

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

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

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* $\equiv (\forall \varphi.\ (\exists x.\ (\varphi)x) \longrightarrow (\exists x.\ \mathit{max}<\varphi>x))$

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

abbreviation *olimitedness* $\equiv (\forall \varphi.\ (\exists x.\ (\varphi)x) \longrightarrow (\exists x.\ \mathit{opt}<\varphi>x))$

abbreviation *osmoothness* $\equiv$
  $(\forall \varphi\ x.\ ((\varphi)x \longrightarrow (\mathit{opt}<\varphi>x \vee (\exists y.\ (y\ \mathbf{r}\ x \wedge \neg(x\ \mathbf{r}\ y) \wedge \mathit{opt}<\varphi>y)))))$

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

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

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

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

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

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

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*: $Suzumura \equiv \forall\, x\ y.\ tcr\ x\ y \longrightarrow \neg\,(y\ \mathbf{r}\ x \wedge \neg\,(x\ \mathbf{r}\ y))$ **by** *simp*

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
by (*simp add*: *assms*(*1*) *assms*(*2*))

We study the relationships between these candidate weakenings of transitivity.

theorem *T3*:
assumes *transitivity*
shows *Suzumura*
— sledgehammer
by (*metis assms sub-rel-def tcr-def transitive-def*)

theorem *T4*:
assumes *transitivity*
shows *Quasitransit*
— sledgehammer
by (*metis assfactor-def assms*)

theorem *T5*:
assumes *Suzumura*
shows *loopfree*
— sledgehammer
by (*metis* (*no-types*, *lifting*) *assms sub-rel-def tcr-def tcr-strict-def*)

theorem *T6*:
assumes *Quasitransit*

shows *loopfree*
— sledgehammer
by (*smt* (*verit*, *best*) *assfactor-def assms sub-rel-def tcr-strict-def transitive-def*)

theorem *T7*:
assumes *reflexivity* **and** *Ferrers*
shows *Quasitransit*
— sledgehammer
by (*metis assfactor-def assms*)

3 Meta-Logical Study

3.1 Correspondence - Max rule

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

lemma *MP*: $[\![\lfloor\varphi\rfloor; \lfloor\varphi \rightarrow \psi\rfloor]\!] \Longrightarrow \lfloor\psi\rfloor$
— sledgehammer
by *simp*

lemma *NEC*: $\lfloor\varphi\rfloor \Longrightarrow \lfloor\Box\varphi\rfloor$
— sledgehammer
by *simp*

$\Box$ is an S5 modality

lemma *C-1-refl*: $\lfloor\Box\varphi \rightarrow \varphi\rfloor$
— sledgehammer
by *simp*

lemma *C-1-trans*: $\lfloor\Box\varphi \rightarrow (\Box(\Box\varphi))\rfloor$
— sledgehammer
by *simp*

lemma *C-1-sym*: $\lfloor\varphi \rightarrow (\Box(\Diamond\varphi))\rfloor$
— sledgehammer
by *simp*

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

lemma *Abs*: $\lfloor\bigcirc<\psi|\varphi> \rightarrow \Box\bigcirc<\psi|\varphi>\rfloor$
— sledgehammer
by *blast*

lemma *Nec*: $\lfloor\Box\psi \rightarrow \bigcirc<\psi|\varphi>\rfloor$
— sledgehammer
by *blast*

lemma *Ext*: $\lfloor\Box(\varphi_1\leftrightarrow\varphi_2) \rightarrow (\bigcirc<\psi|\varphi_1> \leftrightarrow \bigcirc<\psi|\varphi_2>)\rfloor$
— sledgehammer

by *simp*

lemma *Id*: $\lfloor \bigcirc<\varphi|\varphi> \rfloor$
— sledgehammer
by *blast*

lemma *Sh*: $\lfloor \bigcirc<\psi|\varphi_1 \wedge \varphi_2> \rightarrow \bigcirc<(\varphi_2 \rightarrow \psi)|\varphi_1> \rfloor$
— sledgehammer
by *blast*

lemma *COK*:$\lfloor \bigcirc<(\psi_1 \rightarrow \psi_2)|\varphi> \rightarrow (\bigcirc<\psi_1|\varphi> \rightarrow \bigcirc<\psi_2|\varphi>) \rfloor$
— sledgehammer
by *blast*

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

lemma $\lfloor \Diamond\varphi \rightarrow (\bigcirc<\psi|\varphi> \rightarrow P<\psi|\varphi>) \rfloor$
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

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

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

Now we identify a number of correspondences under the max rule. Only the direction property => 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∗*: $\lfloor \Diamond\varphi \rightarrow \bigcirc<\psi|\varphi> \rightarrow P<\psi|\varphi> \rfloor$
— sledgehammer
by (*metis assms*)

lemma
assumes *D∗*: $\lfloor \Diamond\varphi \rightarrow \neg(\bigcirc<\psi|\varphi> \wedge \bigcirc<\neg\psi|\varphi>) \rfloor$
shows *mlimitedness*
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

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

theorem *T9*:
assumes *msmoothness*
shows *CM*: $\lfloor (\bigcirc<\psi|\varphi> \wedge \bigcirc<\chi|\varphi>) \rightarrow \bigcirc<\chi|\varphi \wedge \psi> \rfloor$

— sledgehammer
using *assms* **by** *force*

lemma
assumes *CM*: $\lfloor(\bigcirc<\psi|\varphi> \wedge \bigcirc<\chi|\varphi>) \rightarrow \bigcirc<\chi|\varphi\wedge\psi>\rfloor$
shows *msmoothness*
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

Interval order corresponds to disjunctive rationality, the distinctive axiom of **F**+(DR), but not the other way around.

lemma
assumes *reflexivity*
shows *DR*: $\lfloor\bigcirc<\chi|\varphi\vee\psi> \rightarrow (\bigcirc<\chi|\varphi> \vee \bigcirc<\chi|\psi>)\rfloor$
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

theorem *T10*:
assumes *reflexivity* **and** *Ferrers*
shows *DR*: $\lfloor\bigcirc<\chi|(\varphi\vee\psi)> \rightarrow (\bigcirc<\chi|\varphi> \vee \bigcirc<\chi|\psi>)\rfloor$
— sledgehammer
by (*metis assms*(*1*) *assms*(*2*))

lemma
assumes *DR*: $\lfloor\bigcirc<\chi|\varphi\vee\psi> \rightarrow (\bigcirc<\chi|\varphi> \vee \bigcirc<\chi|\psi>)\rfloor$
shows *reflexivity*
nitpick [*expect=genuine*,*card i=1*] — counterexample found
oops

lemma
assumes *DR*: $\lfloor\bigcirc<\chi|\varphi\vee\psi>\rightarrow(\bigcirc<\chi|\varphi> \vee \bigcirc<\chi|\psi>)\rfloor$
shows *Ferrers*
nitpick [*expect=genuine*,*card i=2*] — counterexample found
oops

Transitivity and totality jointly correspond to the Spohn axiom (Sp), the distinctive axiom of system **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*: $\lfloor(P<\psi|\varphi> \wedge \bigcirc<(\psi\rightarrow\chi)|\varphi>) \rightarrow \bigcirc<\chi|(\varphi\wedge\psi)>\rfloor$
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

lemma
assumes *totality*
shows *Sp*: $\lfloor(P<\psi|\varphi> \wedge \bigcirc<(\psi\rightarrow\chi)|\varphi>) \rightarrow \bigcirc<\chi|(\varphi\wedge\psi)>\rfloor$
nitpick [*expect=genuine*,*card i=3*] — counterexample found

oops

theorem *T11*:
assumes *transitivity* **and** *totality*
shows *Sp*: $\lfloor(P{<}\psi|\varphi{>} \wedge \bigcirc{<}(\psi{\rightarrow}\chi)|\varphi{>}) \rightarrow \bigcirc{<}\chi|(\varphi{\wedge}\psi){>}\rfloor$
— sledgehammer
by (*metis assms*)

theorem *T12*:
assumes *transitivity* **and** *totality*
shows *transit*: $\lfloor(P{<}\varphi|\varphi{\vee}\psi{>} \wedge P{<}\psi|\psi{\vee}\chi{>}) \rightarrow P{<}\varphi|(\varphi{\vee}\chi){>}\rfloor$
— sledgehammer
by (*metis assms*(*1*) *assms*(*2*))

lemma
assumes *Sp*: $\lfloor(P{<}\psi|\varphi{>} \wedge \bigcirc{<}(\psi{\rightarrow}\chi)|\varphi{>}) \rightarrow \bigcirc{<}\chi|(\varphi{\wedge}\psi){>}\rfloor$
shows *totality*
nitpick [*expect=genuine,card i=1*] — counterexample found
oops

lemma
assumes *Sp*: $\lfloor(P{<}\psi|\varphi{>} \wedge \bigcirc{<}(\psi{\rightarrow}\chi)|\varphi{>}) \rightarrow \bigcirc{<}\chi|(\varphi{\wedge}\psi){>}\rfloor$
shows *transitivity*
nitpick [*expect=genuine,card i=2*] — counterexample found
oops

3.2 Correspondence - Opt Rule

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

theorem *T13*:
assumes *olimitedness*
shows *D*: $\lfloor\Diamond\varphi \rightarrow \odot{<}\psi|\varphi{>} \rightarrow \mathcal{P}{<}\psi|\varphi{>}\rfloor$
— sledgehammer
by (*simp add*: *assms*)

lemma
assumes *D*: $\lfloor\Diamond\varphi \rightarrow \odot{<}\psi|\varphi{>} \rightarrow \mathcal{P}{<}\psi|\varphi{>}\rfloor$
shows *olimitedness*
nitpick [*expect=genuine,card i=1*] — counterexample found
oops

Smoothness implies cautious monotony, but not vice-versa.

theorem *T14*:
assumes *osmoothness*
shows *CM'*: $\lfloor(\odot{<}\psi|\varphi{>} \wedge \odot{<}\chi|\varphi{>}) \rightarrow \odot{<}\chi|\varphi{\wedge}\psi{>}\rfloor$
— sledgehammer
using *assms* **by** *force*

lemma

assumes *CM*: $\lfloor(\odot<\psi|\varphi> \wedge \odot<\chi|\varphi>) \rightarrow \odot<\chi|\varphi\wedge\psi>\rfloor$
shows *osmoothness*
nitpick [*expect=genuine,card i=1*] — counterexample found
oops

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

theorem *T15*:
assumes *transitivity*
shows *Sp'*: $\lfloor(\mathcal{P}<\psi|\varphi> \wedge \odot<(\psi\rightarrow\chi)|\varphi>) \rightarrow \odot<\chi|(\varphi\wedge\psi)>\rfloor$
— sledgehammer
by (*metis assms*)

theorem *T16*:
assumes *transitivity*
shows *Trans'*: $\lfloor(\mathcal{P}<\varphi|\varphi\vee\psi> \wedge \mathcal{P}<\psi|\psi\vee\xi>)\rightarrow \mathcal{P}<\varphi|\varphi\vee\xi>\rfloor$
— sledgehammer
by (*metis assms*)

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

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

lemma
assumes *reflexivity*
shows *DR'*: $\lfloor\odot<\chi|\varphi\vee\psi> \rightarrow (\odot<\chi|\varphi> \vee \odot<\chi|\psi>)\rfloor$
nitpick [*expect=genuine,card i=3*] — counterexample found
oops

theorem *T17*:
assumes *reflexivity* **and** *Ferrers*
shows *DR'*: $\lfloor\odot<\chi|\varphi\vee\psi> \rightarrow (\odot<\chi|\varphi> \vee \odot<\chi|\psi>)\rfloor$
— sledgehammer
by (*metis assms*(*2*))

lemma
assumes *DR*: $\lfloor\odot<\chi|\varphi\vee\psi> \rightarrow (\odot<\chi|\varphi> \vee \odot<\chi|\psi>)\rfloor$
shows *reflexivity*
nitpick [*expect=genuine,card i=1*] — counterexample found
oops

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

oops

3.3 Correspondence - Lewis' rule

We have deontic explosion under the max rule.

theorem *DEX*: $\lfloor(\diamond\varphi \wedge \bigcirc{<}\psi|\varphi{>} \wedge \bigcirc{<}\neg\psi|\varphi{>}) \rightarrow \bigcirc{<}\chi|\varphi{>}\rfloor$
— sledgehammer
by *blast*

But no deontic explosion under Lewis' rule.

lemma *DEX*: $\lfloor(\diamond\varphi \wedge \circ{<}\psi|\varphi{>} \wedge \circ{<}\neg\psi|\varphi{>}) \rightarrow \circ{<}\chi|\varphi{>}\rfloor$
nitpick [*expect=genuine*,*card i=2*] — counterexample found
oops

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

theorem *T18*:
assumes *mlimitedness* **and** *transitivity* **and** *totality*
shows $\lfloor\circ{<}\psi|\varphi{>}\leftrightarrow\odot{<}\psi|\varphi{>}\rfloor$
— sledgehammer
by (*smt* (*z3*) *assms*)

theorem *T19*:
assumes *mlimitedness* **and** *transitivity* **and** *totality*
shows $\lfloor\circ{<}\psi|\varphi{>}\leftrightarrow\bigcirc{<}\psi|\varphi{>}\rfloor$
— sledgehammer
by (*smt* (*z3*) *assms*)

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

theorem *Abs'*: $\lfloor\circ{<}\psi|\varphi{>} \rightarrow \Box\circ{<}\psi|\varphi{>}\rfloor$
— sledgehammer
by *auto*

theorem *Nec'*: $\lfloor\Box\psi \rightarrow \circ{<}\psi|\varphi{>}\rfloor$
— sledgehammer
by *auto*

theorem *Ext'*: $\lfloor\Box(\varphi_1\leftrightarrow\varphi_2) \rightarrow (\circ{<}\psi|\varphi_1{>} \leftrightarrow \circ{<}\psi|\varphi_2{>})\rfloor$
— sledgehammer
by *auto*

theorem *Id'*: $\lfloor\circ{<}\varphi|\varphi{>}\rfloor$
— sledgehammer
by *auto*

theorem *Sh'*: $\lfloor\circ{<}\psi|\varphi_1\wedge\varphi_2{>} \rightarrow \circ{<}(\varphi_2\rightarrow\psi)|\varphi_1{>}\rfloor$
— sledgehammer
by *auto*

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*: $\lfloor\Diamond\varphi \rightarrow (\circ<\psi|\varphi> \rightarrow \int<\psi|\varphi>)\rfloor$
 nitpick [*expect=genuine*,*card i=2*] — counterexample found
 oops

lemma *Sp*: $\lfloor(\int<\psi|\varphi> \wedge \circ<(\psi\rightarrow\chi)|\varphi>) \rightarrow \circ<\chi|(\varphi\wedge\psi)>\rfloor$
 nitpick [*expect=genuine*,*card i=3*] — counterexample found
 oops

lemma *COK*:$\lfloor\circ<(\psi_1\rightarrow\psi_2)|\varphi> \rightarrow (\circ<\psi_1|\varphi> \rightarrow \circ<\psi_2|\varphi>)\rfloor$
 nitpick [*expect=genuine*,*card i=2*] — counterexample found
 oops

lemma *CM*: $\lfloor(\circ<\psi|\varphi>\wedge\circ<\chi|\varphi>)\rightarrow \circ<\chi|\varphi\wedge\psi>\rfloor$
 nitpick [*expect=genuine*,*card i=2*] — counterexample found
 oops

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

theorem *T20*:
 assumes *totality*
 shows $\lfloor\Diamond\varphi \rightarrow (\circ<\psi|\varphi> \rightarrow \int<\psi|\varphi>)\rfloor$
 — sledgehammer
 using *assms* **by** *blast*

lemma
 assumes $\lfloor\Diamond\varphi \rightarrow (\circ<\psi|\varphi> \rightarrow \int<\psi|\varphi>)\rfloor$
 shows *totality*
 nitpick [*expect=genuine*,*card i=3*] — counterexample found
 oops

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

theorem *T21*:
 assumes *transitivity*
 shows *Sp''*: $\lfloor(\int<\psi|\varphi> \wedge \circ<(\psi\rightarrow\chi)|\varphi>) \rightarrow \circ<\chi|(\varphi\wedge\psi)>\rfloor$
 — sledgehammer
 using *assms* **by** *blast*

theorem *T22*:
 assumes *transitivity*
 shows *Tr''*: $\lfloor(\int<\varphi|\varphi\vee\psi>\wedge\int<\psi|\psi\vee\chi>)\rightarrow \int<\varphi|\varphi\vee\chi>\rfloor$
 — sledgehammer
 using *assms* **by** *blast*

lemma
 assumes *Sp''*: $\lfloor(\int<\psi|\varphi> \wedge \circ<(\psi\rightarrow\chi)|\varphi>) \rightarrow \circ<\chi|(\varphi\wedge\psi)>\rfloor$
 shows *transitivity*
 nitpick — counterexample found

oops

lemma
assumes *Tr''*: $\lfloor(\int <\varphi|\varphi\vee\psi>\wedge\int <\psi|\psi\vee\chi>)\rightarrow \int <\varphi|\varphi\vee\chi>\rfloor$
shows *transitivity*
nitpick — counterexample found
oops

lemma
assumes *transitivity*
shows *COK*:$\lfloor\circ<(\psi_1\rightarrow\psi_2)|\varphi> \rightarrow (\circ<\psi_1|\varphi> \rightarrow \circ<\psi_2|\varphi>)\rfloor$
nitpick [*expect=genuine*,*card i=2*] — counterexample found
oops

lemma
assumes *totality*
shows *COK*:$\lfloor\circ<(\psi_1\rightarrow\psi_2)|\varphi> \rightarrow (\circ<\psi_1|\varphi> \rightarrow \circ<\psi_2|\varphi>)\rfloor$
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

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'*:$\lfloor\circ<(\psi_1\rightarrow\psi_2)|\varphi> \rightarrow (\circ<\psi_1|\varphi> \rightarrow \circ<\psi_2|\varphi>)\rfloor$
— sledgehammer
by (*smt* (*verit*, *ccfv-SIG*) *assms*(*1*) *assms*(*2*))

lemma
assumes *COK'*:$\lfloor\circ<(\psi_1\rightarrow\psi_2)|\varphi> \rightarrow (\circ<\psi_1|\varphi> \rightarrow \circ<\psi_2|\varphi>)\rfloor$
shows *transitivity* **and** *totality*
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

theorem *T24*:
assumes *transitivity* **and** *totality*
shows *CM''*: $\lfloor(\circ<\psi|\varphi>\wedge\circ<\chi|\varphi>)\rightarrow \circ<\chi|\varphi\wedge\psi>\rfloor$
— sledgehammer
by (*metis assms*)

lemma
assumes *CM''*: $\lfloor(\circ<\psi|\varphi>\wedge\circ<\chi|\varphi>)\rightarrow \circ<\chi|\varphi\wedge\psi>\rfloor$
shows *transitivity* **and** *totality*
nitpick [*expect=genuine*,*card i=3*] — counterexample found
oops

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

theorem *T25*:
assumes *transitivity*

shows $\lfloor(\mathcal{P}<\psi|\varphi> \wedge \odot<(\psi\rightarrow\chi)|\varphi>) \rightarrow \odot<\chi|(\varphi\wedge\psi)>\rfloor$
— sledgehammer
by (*metis assms*)

lemma
assumes *transitivity*
shows $\lfloor(\mathcal{P}<\varphi|\varphi\vee\psi> \wedge \mathcal{P}<\xi|\psi\vee\xi>)\rightarrow\mathcal{P}<\xi|\varphi\vee\xi>\rfloor$
nitpick [*expect=genuine,card i=2*] — counterexample found
oops

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

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 $A::\sigma$ $Aplus::\sigma$ $B::\sigma$

Here is the formalization of the paradox.

axiomatization where
— A is striclty better than B
P0: $\lfloor(\neg\odot<\neg A|A\vee B>\wedge\odot<\neg B|A\vee B>)\rfloor$ **and**
— Aplus is at least as good as A
P1: $\lfloor\neg\odot<\neg Aplus|A\vee Aplus>\rfloor$ **and**
— B is strictly better than Aplus
P2: $\lfloor(\neg\odot<\neg B|Aplus\vee B> \wedge \odot<\neg Aplus|Aplus\vee B>)\rfloor$

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

theorem *T0*:
assumes *transitivity*

shows *False*
— sledgehammer
by (*metis P0 P1 P2 assms*)

Nitpick shows consistency in the absence of transitivity:

theorem *T1*:
True
nitpick [*satisfy*,*expect*=*genuine*,*card i*=*3*] — model found
oops

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
by (*metis P0 P1 P2 assms*(*2*))

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
oops

theorem *T4*:
assumes *Quasitransit*
shows *True*
nitpick [*satisfy*,*expect*=*genuine*,*card*=*4*] — model found
oops

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

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

theorem *T0*:
assumes *transitivity*
shows *False*
— sledgehammer
by (*metis PPP0 PPP1 PPP2 assms*)

Nitpick shows consistency in the absence of transitivity:

lemma *T1*:
True
nitpick [*satisfy*,*expect=genuine*,*card i=3*,*show-all*] — model found
oops

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

theorem *T2*:
assumes *reflexivity Ferrers*
shows *False*
— sledgehammer
by (*metis PPP0 PPP1 PPP2 assms*(*1*) *assms*(*2*))

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
oops

theorem *T4*:
assumes *Quasitransit*
shows *True*
nitpick [*satisfy*,*expect=genuine*,*card=4*] — model found
oops

end

6 The Mere Addition Paradox: Max Rule

There are surprising results with the max rule. Transitivity alone generates an inconsistencty 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 striclty better than B
PP0: $\lfloor(\neg\bigcirc<\neg A|A\vee B>\wedge\bigcirc<\neg B|A\vee B>)\rfloor$ **and**
— Aplus is at least as good as A
PP1: $\lfloor\neg\bigcirc<\neg Aplus|A\vee Aplus>\rfloor$ **and**
— B is strictly better than Aplus
PP2: $\lfloor(\neg\bigcirc<\neg B|Aplus\vee B> \wedge \bigcirc<\neg Aplus|Aplus\vee B>)\rfloor$

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
  oops

Nitpick shows consistency in the absence of transitivity:

theorem *T1*:
  shows *True*
  nitpick [*satisfy*,*expect*=*genuine*,*card i*=*3*] — model found
  oops

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

theorem *T2*:
  assumes *reflexivity* **and** *Ferrers*
  shows *False*
  — sledgehammer
  by (*metis PP0 PP1 PP2 assms*(*1*) *assms*(*2*))

Nitpick shows consistency if transitivity is weakened into acyclicity:

theorem *T3*:
  assumes *loopfree*
  shows *True*

nitpick [*satisfy*,*expect=genuine*,*card=3*] — model found
oops

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: $\forall y.\ y{=}i1 \lor y{=}i2 \lor y{=}i3 \lor y{=}i4 \lor y{=}\ i5 \lor y{=}\ i6 \lor y{=}\ i7$
shows *False*
using *assfactor-def PP0 PP1 PP2 assms*
— sledgehammer()
— proof found by Sledgehammer, but reconstruction fails
oops

theorem *T5*:
assumes
Quasitransit **and**
OnlyOnes: $\forall y.\ y{=}i1 \lor y{=}i2 \lor y{=}i3 \lor y{=}i4 \lor y{=}\ i5 \lor y{=}\ i6 \lor y{=}i7$
shows *False*
using *assfactor-def PP0 PP1 PP2 assms*
— sledgehammer()
— proof found by Sledgehammer, but reconstruction fails
oops

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* $\equiv \exists M.\ (\exists z{::}i.\ \neg(M\ z) \land (\exists G.\ (\forall y{::}i.\ (\exists x.\ (M\ x) \land (G\ x) = y))))$

lemma *infinity* **nitpick**[*expect=genuine*] **oops** — countermodel found

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
oops

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
oops

Transitivity and totality together give inconsistency:

theorem *T0′*:
assumes *transitivity* **and** *totality*
shows *False*
— sledgehammer
by (*metis PP0 PP1 PP2 assms*(*1*) *assms*(*2*))

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.