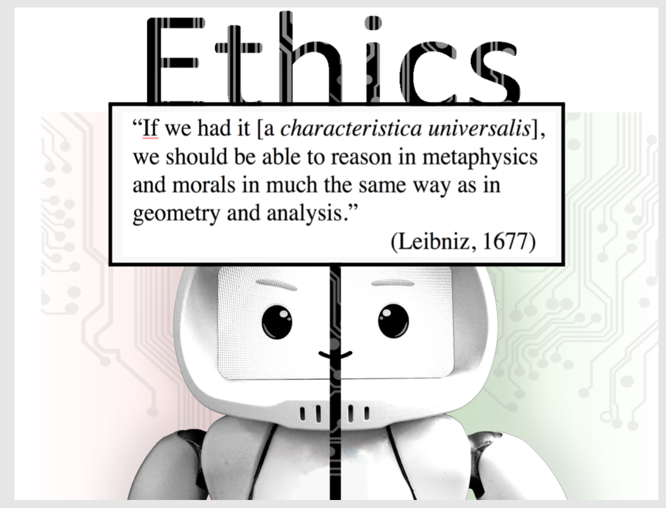
Studies in Computational Metaphysics & Computational (Pseudo-)Ethics

Christoph Benzmüller

Freie Universität Berlin | University of Luxembourg

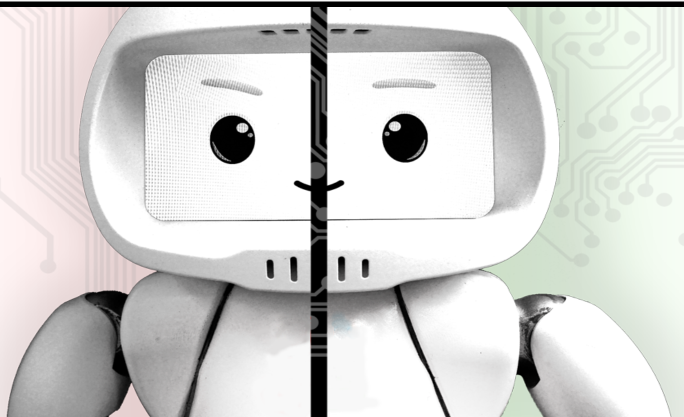


Colloquium Cognitive Systems, U Ulm, 29 Nov 2018

Ethics

Peaceful coexistence with **intelligent autonomous systems (IASs)**?

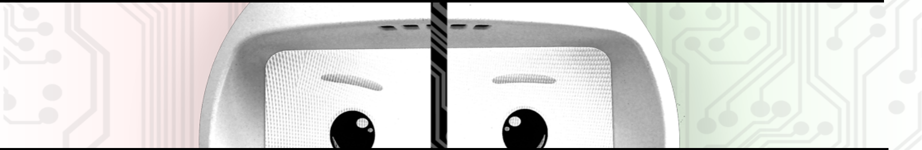
- appropriate forms of **machine-control**
- appropriate forms of **human-machine-interaction**



Ethics

Peaceful coexistence with **intelligent autonomous systems (IASs)**?

- appropriate forms of **machine-control**
- appropriate forms of **human-machine-interaction**



Existing societal processes are based on:

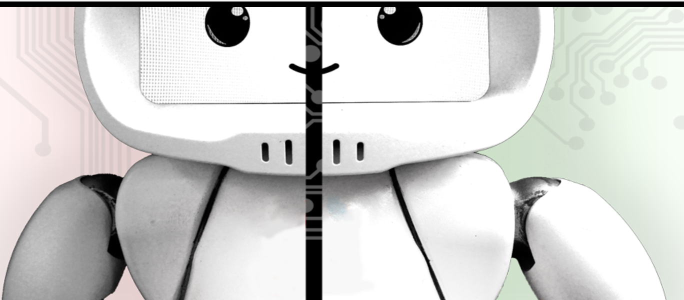
- **rational argumentation & dialog**
- **explicit normative reasoning** (legal & ethical)

Deployment of IASs lacking such competencies? How wise is this?

Ethics

Talk Outline

- **A** **Motivation:** Explicit Ethical Reasoning
- **B** **Technology:** Universal Reasoning in Higher-Order Logic (HOL)
- **C** **Evidence:** Analysis of Rational Arguments in Metaphysics
- **D** **Demo(s):** Normative Reasoning Experimentation Platform



Motivation

Long-term: Emerging Superintelligence Really? Anyhow ...

- How to prevent Superintelligence from turning against humanity?

Medium-term: Development of pseudo-ethical skills in IASs

- Which norms? Which reasoning principles?
- What architectural design? What functionalities?
- How to implement, deploy and verify?

Different kinds of systems and approaches:

- [Moor, 2009]:
  - ethical impact agents (ethical consequences to actions)
  - implicit ethical agents (ethical reactions to given situations)
  - explicit ethical agents (reasoning with ethical theories/rules)
  - full ethical agents (conscious, intentional, free will)
- bottom-up vs. top-down
- [DoranEtAl., 2017]:
  opaque — comprehensible — interpretable — explainable AI

Motivation

Long-term: Emerging Superintelligence Really? Anyhow . . .

- How to prevent Superintelligence from turning against humanity?

Medium-term: Development of pseudo-ethical skills in IASs

- Which norms? Which reasoning principles?
- What architectural design? What functionalities?
- How to implement, deploy and verify?

Different kinds of systems and approaches:

- [Moor, 2009]:
  – ethical impact agents (ethical consequences to actions)
  – implicit ethical agents (ethical reactions to given situations)
  – explicit ethical agents (reasoning with ethical theories/rules)
  – full ethical agents (conscious, intentional, free will)
- bottom-up vs. top-down
- [DoranEtAl., 2017]:
  opaque — comprehensible — interpretable — explainable AI

Long-term: Emerging Superintelligence Really? Anyhow . . .

- How to prevent Superintelligence from turning against humanity?

Medium-term: Development of pseudo-ethical skills in IASs

- Which norms? Which reasoning principles?
- What architectural design? What functionalities?
- How to implement, deploy and verify?

Different kinds of systems and approaches:

- [Moor, 2009]:
  – ethical impact agents (ethical consequences to actions)
  – implicit ethical agents (ethical reactions to given situations)
  – explicit ethical agents (reasoning with ethical theories/rules)
  – full ethical agents (conscious, intentional, free will)
- bottom-up vs. top-down
- [DoranEtAl., 2017]:
  opaque — comprehensible — interpretable — explainable AI

Motivation

Long-term: Emerging Superintelligence Really? Anyhow . . .

- How to prevent Superintelligence from turning against humanity?

Medium-term: Development of pseudo-ethical skills in IASs

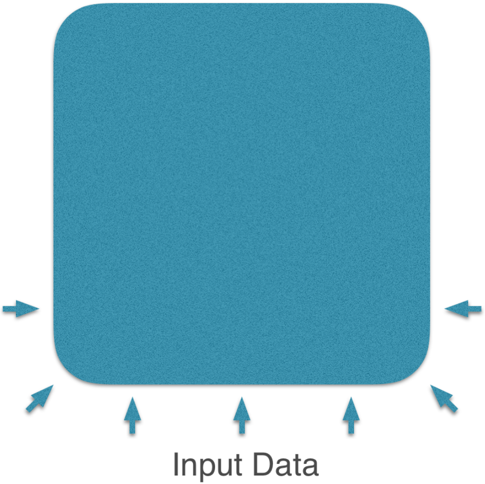
- Which norms? Which reasoning principles?
- What architectural design? What functionalities?
- How to implement, deploy and verify?

Different kinds of systems and approaches:

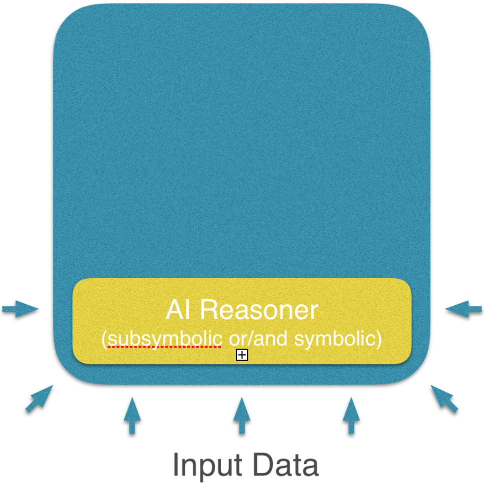
- [Moor, 2009]:
  – ethical impact agents (ethical consequences to actions)
  – implicit ethical agents (ethical reactions to given situations)
  – explicit ethical agents (reasoning with ethical theories/rules)
  – full ethical agents (conscious, intentional, free will)
- bottom-up vs. top-down
- [DoranEtAl., 2017]:
  opaque — comprehensible — interpretable — explainable AI

Pseudo-Ethical IAS (medium-term)

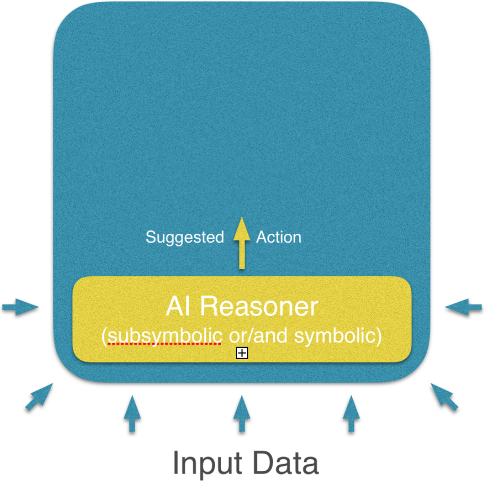
Pseudo-Ethical IAS (medium-term)



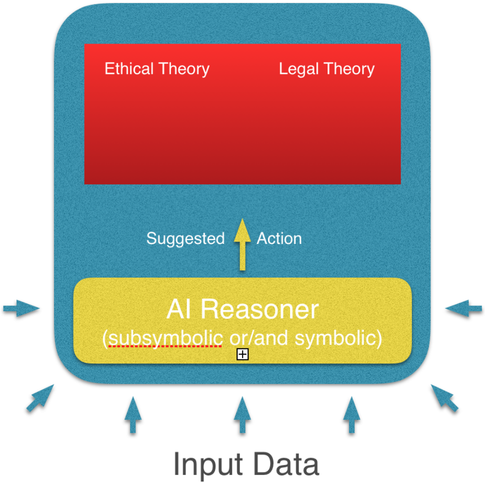
Pseudo-Ethical IAS (medium-term)



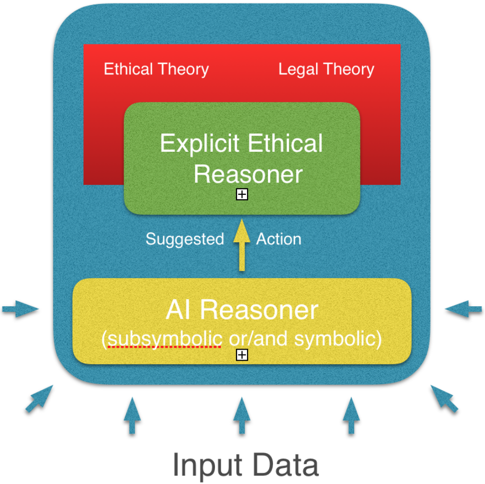
Pseudo-Ethical IAS (medium-term)



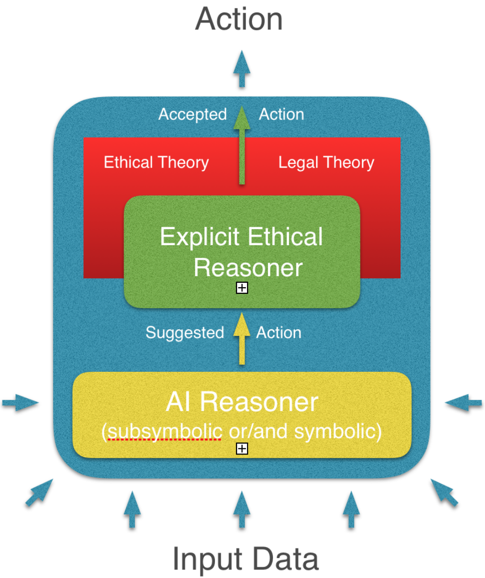
Pseudo-Ethical IAS (medium-term)



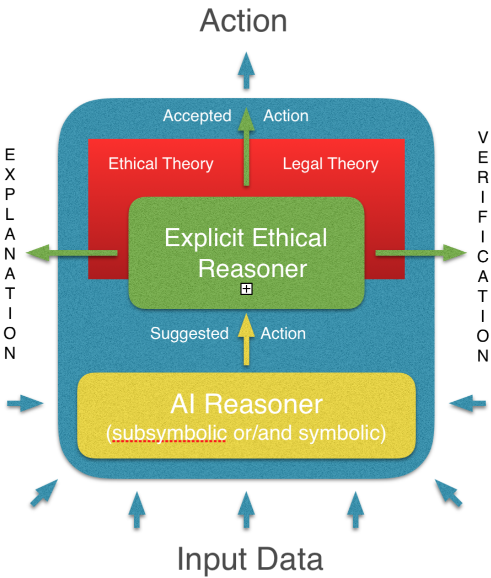
Pseudo-Ethical IAS (medium-term)



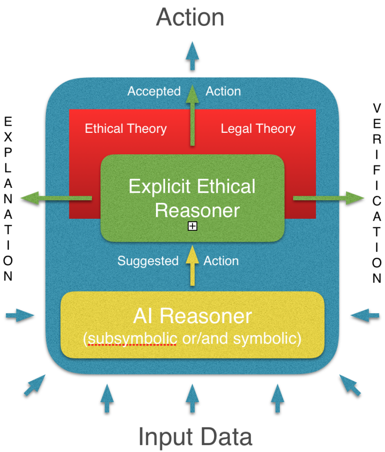
Pseudo-Ethical IAS (medium-term)



Pseudo-Ethical IAS (medium-term)



Pseudo-Ethical IAS (medium-term)



Related Work

- Artificial Moral Agents
  - [Wallach&Allen, 2008]
- Ethical Governors
  - [ArkinEtAl., 2009, 2012]
  - [Dennis&Fisher, 2017]
- Ethical Deliberation in ART
  - [Dignum, 2017]
- Programming Machine Ethics
  - [Pereira&Saptawijaya, 2016]
- . . .

Strategical Relevance of Research Direction

Bundesregierung (Nov 2018): Strategie Künstliche Intelligenz

"Ethische und rechtsstaatliche Anforderungen sollten als integraler Bestandteil — und damit Markenzeichen einer „AI made in Europe" — im gesamten Prozess der Entwicklung und Anwendung von KI Beachtung finden. Dies umfasst die Forschung, Entwicklung und die Produktion von KI, aber auch den Einsatz, den Betrieb, die Kontrolle und die Governance KI-basierter Anwendungen. Entwicklung von Verfahren zur Kontrolle und Nachvollziehbarkeit algorithmischer Entscheidungen sollte alle Akteure, inkl. Industrie, einbeziehen."

`https://www.bmbf.de/files/Nationale_KI-Strategie.pdf`; page 40

Ben Goertzel (CEO SingularityNET; Nov 2018): "Toward Democratic, Lawful Citizenship for AIs, Robots, and Corporations"

"Being an effective citizen of a nation operating under rule of law requires a form of general intelligence that combines formal linguistic and symbolic knowledge (the legal code) with the ability to abstract patterns from multimodal sensory data and informal linguistic data (corresponding to actual real-life situations to which the law needs to be applied). So an AI Citizenship Test needs to be a particular form of a General Intelligence Test. And it needs to be a test that stresses one of the most interesting issues at the core of modern AI R&D: the fusion of symbolic and subsymbolic knowledge."

`https://tinyurl.com/y8h94ouv`

Which Reasoning Formalisms?

"If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis."

(Leibniz, 1677)

Challenges for Explicit Ethical Reasoning Engines: Which Logic(s)?

- Dilemmas, conflicting theories, etc.
- Appropriate modeling-of/reasoning-with notion of **obligation**
  - Contrary-to-duty (**CTD**) scenarios

Which Reasoning Formalisms?

"If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis."
(Leibniz, 1677)

Challenges for Explicit Ethical Reasoning Engines: Which Logic(s)?

- Dilemmas, conflicting theories, etc.
- Appropriate modeling-of/reasoning-with notion of **obligation**
  - Contrary-to-duty (**CTD**) scenarios

Standard CTD structure (Chisholm)

1. obligatory 'a'
2. obligatory 'if a then not b'
3. if 'not a' then obligatory 'b'
4. 'not a' (in a given situation)

Danger: Paradox/inconsistency — ex falso quodlibet!

Which Reasoning Formalisms?

> "If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis."
>
> (Leibniz, 1677)

Challenges for Explicit Ethical Reasoning Engines: Which Logic(s)?

- Dilemmas, conflicting theories, etc.
- Appropriate modeling-of/reasoning-with notion of **obligation**
  - Contrary-to-duty (**CTD**) scenarios

CTD example (X. Parent): **EU General Data Protection Regulation (GDPR)**

1. Personal data shall be processed lawfully. (Art. 5)
   E.g., the data subject must have given consent to the processing. (Art. 6/1.a)
2. **Implicit:** The data shall be kept, for the agreed purposes, if processed lawfully.
3. If personal data has been processed unlawfully, the controller has the obligation to erase the personal data in question without delay. (Art. 17.d, right to be forgotten)
4. **Given situation:** Some personal data has been processed unlawfully.

Danger: Paradox/inconsistency — ex falso quodlibet!

Which Reasoning Formalisms?

"If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis."

(Leibniz, 1677)

Challenges for Explicit Ethical Reasoning Engines: Which Logic(s)?

- Dilemmas, conflicting theories, etc.
- Appropriate modeling-of/reasoning-with notion of **obligation**
  - Contrary-to-duty (**CTD**) scenarios

L. van der Torre

X. Parent

A. Farjami

Deontic Logic

- Reasoning about obligations and permissions
- Two groups of approaches:

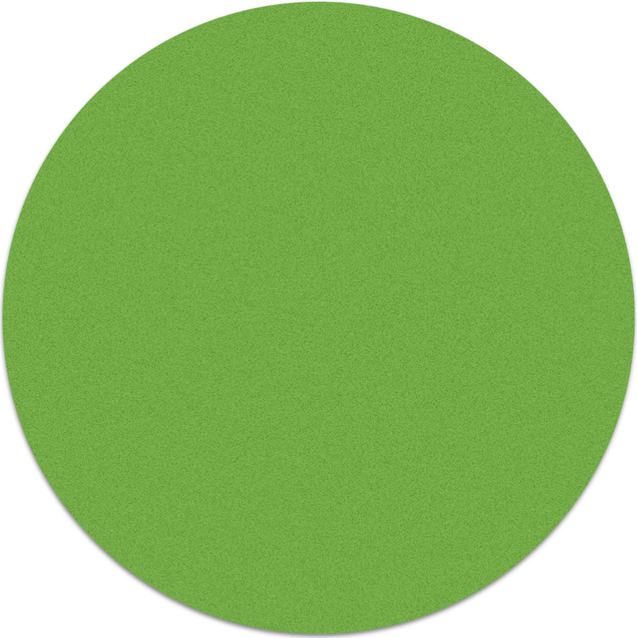
  — Possible worlds
  - standard deontic logic **CTD: no**
  - dyadic deontic logic **CTD: yes**

  — Norm-based semantics
  - input/output logic **CTD: yes**

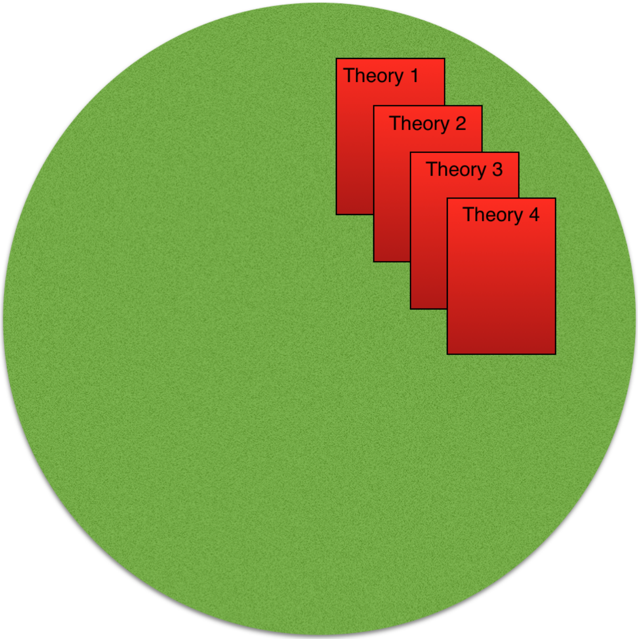
Further interests and challenges

- Combination with other logics (other modalities)
- Propositional deontic logic(s) will hardly be sufficient in practice

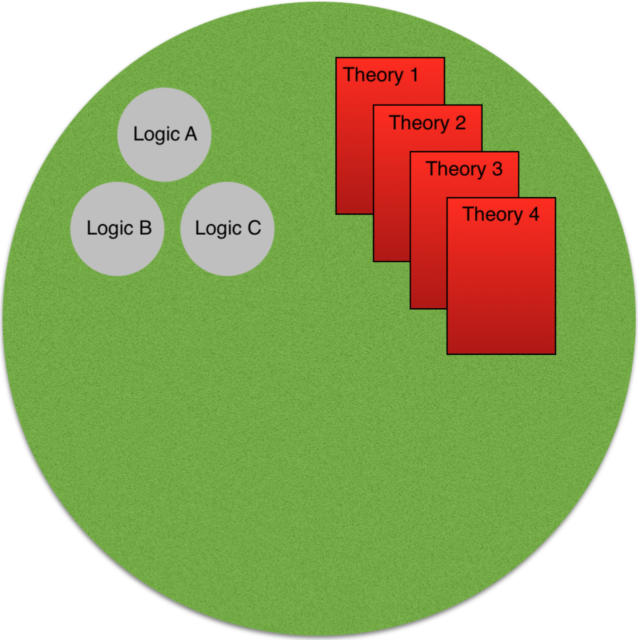
Normative Reasoning Experimentation Platform



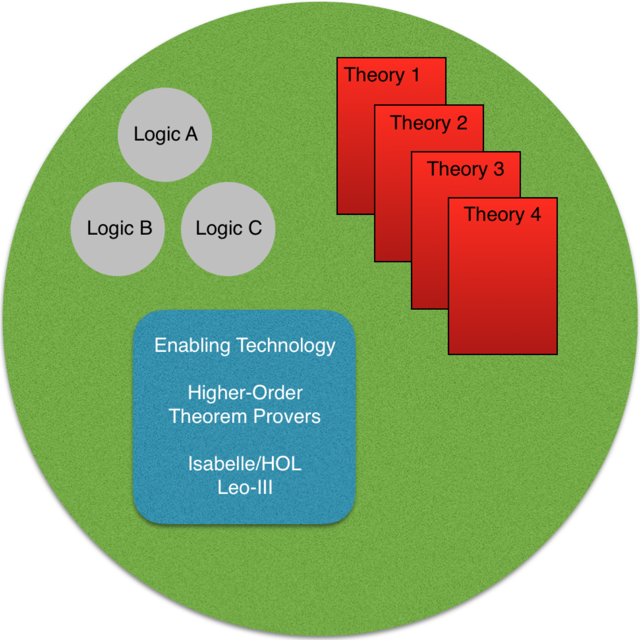
Normative Reasoning Experimentation Platform



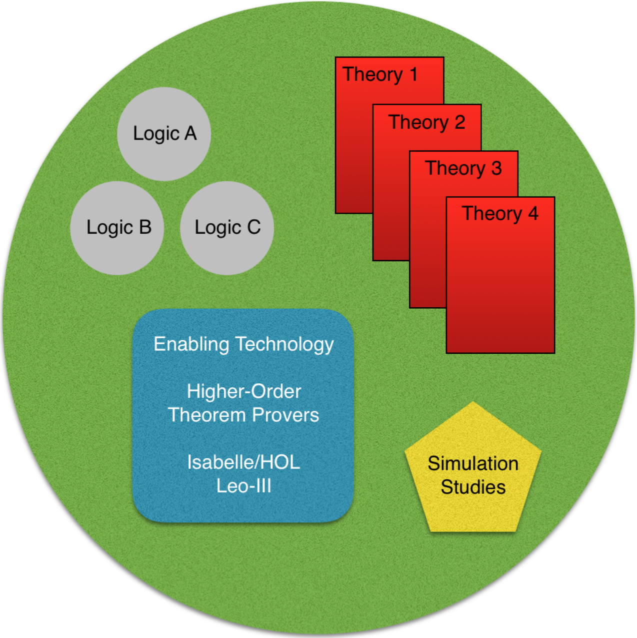
Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform — Demo in Isabelle/HOL

GDPR.thy

GDPR.thy (~/chris/trunk/tex/talks/2018-DEON/DEMO/)

```
theory GDPR imports SDL                      (* Christoph Benzmüller & Xavier Parent, 2018 *)

begin (*** GDPR Example ***)
 consts process_data_lawfully::σ erase_data::σ kill_boss::σ

 axiomatization where
  (* It is an obligation to process data lawfully. *)
    A1: "⌊O⟨process_data_lawfully⟩⌋"  and
  (* Implicit: It is an obligation to keep the data if it was processed lawfully. *)
    Implicit: "⌊O⟨process_data_lawfully → ¬erase_data⟩⌋" and
  (* If data was not processed lawfully, then it is an obligation to erase the data. *)
    A2: "⌊¬process_data_lawfully → O⟨erase_data⟩⌋"
  (* Given a situation where data is processed unlawfully. *) and
    A3: "⌊¬process_data_lawfully⌋cw"

 (*** Some Experiments ***)
  lemma True nitpick [satisfy] oops  (* Consistency-check: Is there a model? *)
  lemma False sledgehammer       oops  (* Inconsistency-check: Can Falsum be derived? *)

  lemma "⌊O⟨erase_data⟩⌋"   sledgehammer nitpick oops  (* Should the data be erased? *)
  lemma "⌊O⟨¬erase_data⟩⌋"  sledgehammer nitpick oops  (* Should the data be kept? *)
  lemma "⌊O⟨kill_boss⟩⌋"    sledgehammer nitpick oops  (* Should the boss be killed? *)
end
```

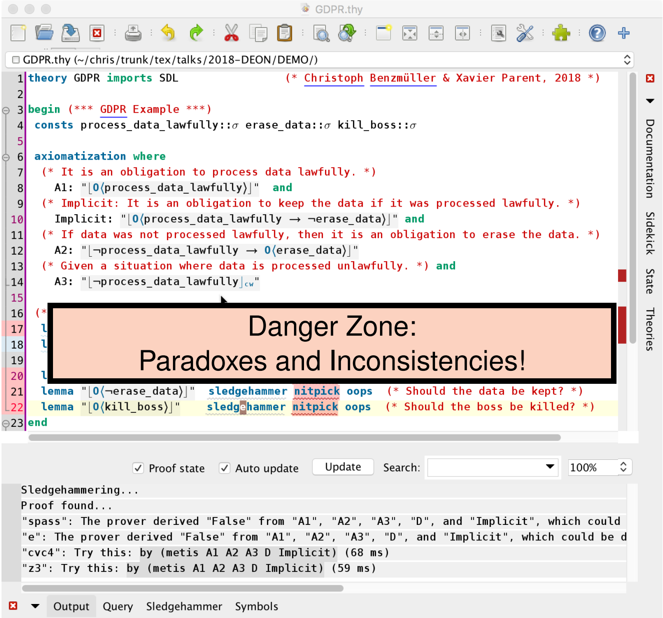
Documentation | Sidekick | State | Theories

Proof state | Auto update | Update | Search: | 100%

```
Sledgehammering...
Proof found...
"spass": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could
"e": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could be d
"cvc4": Try this: by (metis A1 A2 A3 D Implicit) (68 ms)
"z3": Try this: by (metis A1 A2 A3 D Implicit) (59 ms)
```

Output | Query | Sledgehammer | Symbols

Normative Reasoning Experimentation Platform — Demo in Isabelle/HOL

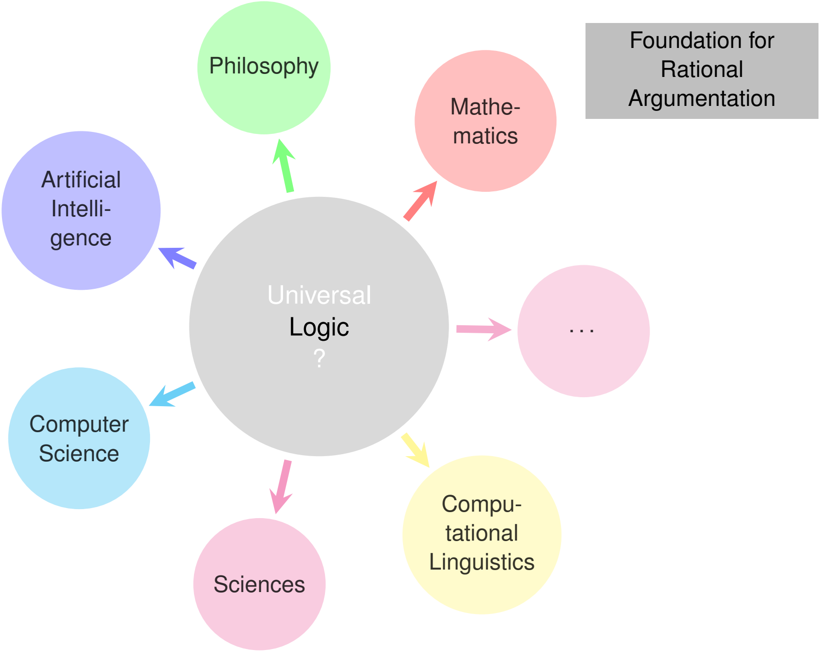


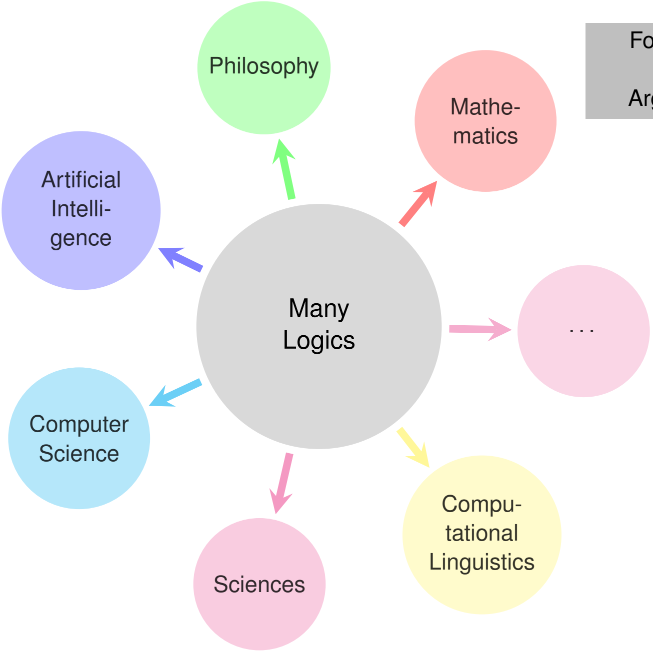
"If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis."

(Leibniz, 1677)

Part B

Technology: Universal Reasoning in Higher-Order Logic


Philosophy
Mathe-
matics
Foundation for
Rational
Argumentation
Artificial
Intelli-
gence
Universal
Logic
?
...
Computer
Science
Compu-
tational
Linguistics
Sciences



Foundation for
Rational
Argumentation

Logic Zoo

Logic Zoo

Classical Logic, of order

0. Propositional Logic
1. First-order Logic
2. Second-order Logic

...

n. Higher-order Logic

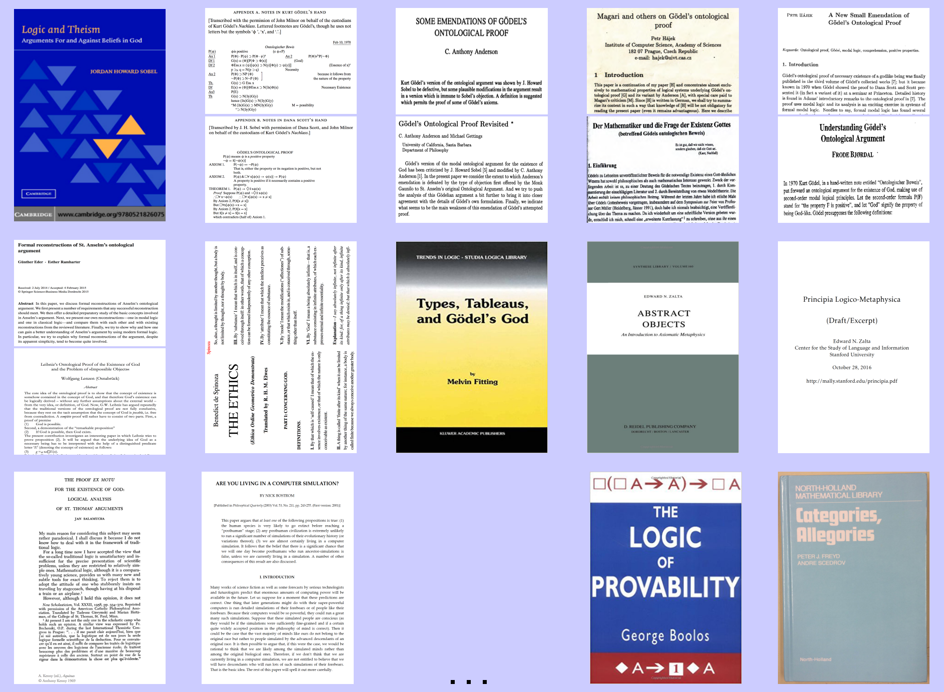
Non-Classical Logics

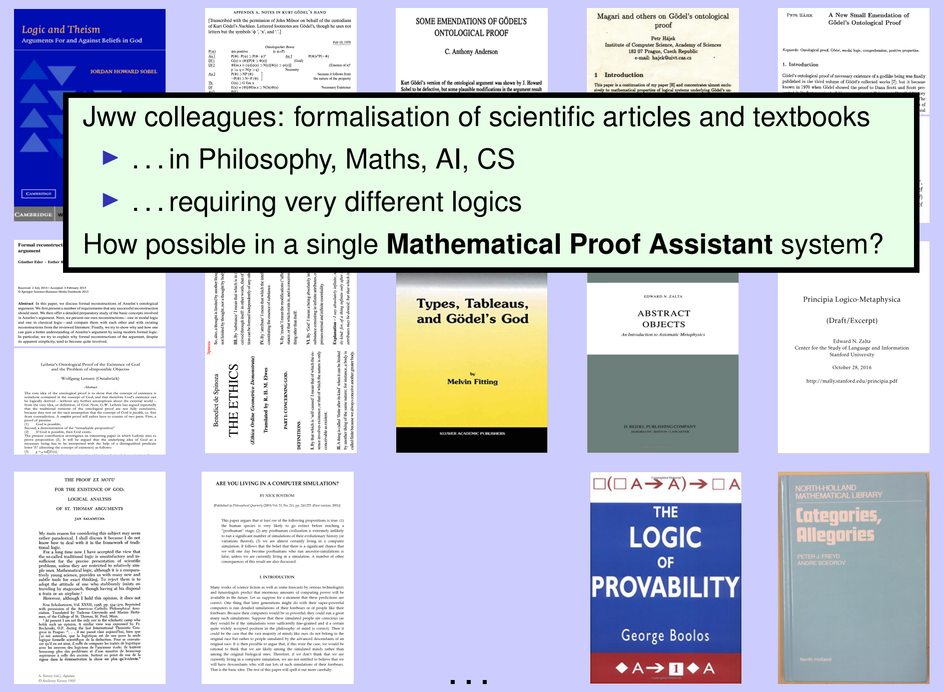
- Intuitionistic/Constructive Logics (incl. Univalent Foundations)
- Modal Logics, Conditional Logics, Temporal Logis, Spatial Logics
- Many-valued Logics
- Paraconsistent Logics
- Free Logics, Inclusive Logics
- Logics for special applications: Ethics, Social Choice, Legal Reasoning, ...
- Separation Logic, ...

Example Application in Metaphysics/Philosophy:

Necessarily, God exists: $\Box \exists x. Gx$

Kurt Gödel's definition of God: $Gx \;:=\; \forall \Phi. Positive\ \Phi \rightarrow \Phi x$


Logic and Theism
Arguments For and Against Beliefs in God
JORDAN HOWARD SOBEL
CAMBRIDGE
www.cambridge.org/9780521826075
APPENDIX A. NOTES IN KURT GÖDEL'S HAND
APPENDIX B. NOTES IN DANA SCOTT'S HAND
SOME EMENDATIONS OF GÖDEL'S ONTOLOGICAL PROOF
C. Anthony Anderson
Gödel's Ontological Proof Revisited
C. Anthony Anderson and Michael Gettings
Magari and others on Gödel's ontological proof
Petr Hájek
1 Introduction
Der Mathematiker und die Frage der Existenz Gottes
A New Small Emendation of Gödel's Ontological Proof
1. Introduction
Understanding Gödel's Ontological Argument
FRODE BJØRDAL
Formal reconstructions of St. Anselm's ontological argument
Günther Eder · Esther Ramharter
Leibniz's Ontological Proof of the Existence of God and the Problem of »Impossible Objects«
Wolfgang Lenzen (Osnabrück)
Benedict de Spinoza
THE ETHICS
(Ethica Ordine Geometrico Demonstrata)
Translated by R. H. M. Elwes
PART I: CONCERNING GOD.
TRENDS IN LOGIC · STUDIA LOGICA LIBRARY
Types, Tableaus, and Gödel's God
by Melvin Fitting
KLUWER ACADEMIC PUBLISHERS
EDWARD N. ZALTA
ABSTRACT OBJECTS
An Introduction to Axiomatic Metaphysics
D. REIDEL PUBLISHING COMPANY
Principia Logico-Metaphysica
(Draft/Excerpt)
Edward N. Zalta
Center for the Study of Language and Information
Stanford University
October 28, 2016
http://mally.stanford.edu/principia.pdf
THE PROOF EX MOTU FOR THE EXISTENCE OF GOD: LOGICAL ANALYSIS OF ST. THOMAS' ARGUMENTS
JAN SALAMUCHA
ARE YOU LIVING IN A COMPUTER SIMULATION?
BY NICK BOSTROM
I. INTRODUCTION
THE LOGIC OF PROVABILITY
George Boolos
NORTH-HOLLAND MATHEMATICAL LIBRARY
Categories, Allegories
PETER J. FREYD
ANDRE SCEDROV
North-Holland

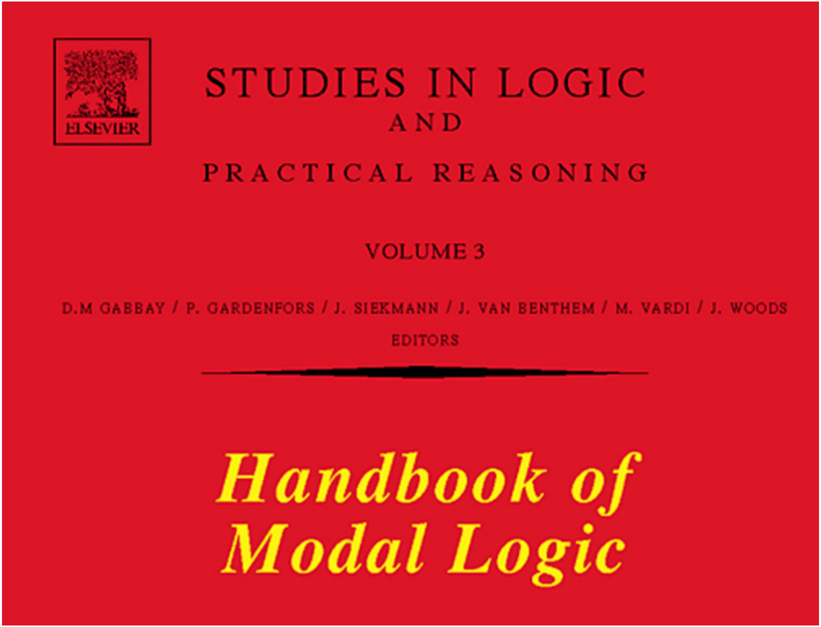


Jww colleagues: formalisation of scientific articles and textbooks

- …in Philosophy, Maths, AI, CS
- …requiring very different logics

How possible in a single **Mathematical Proof Assistant** system?

Example: Modal Logic Textbook



Example: Modal Logic Textbook

2 BASIC MODAL LOGIC

In this section we introduce the basic modal language and its relational semantics. We define basic modal syntax, introduce models and frames, and give the satisfaction definition. We then draw the reader's attention to the internal perspective that modal languages offer on relational structure, and explain why models and frames should be thought of as graphs. Following this we give the standard translation. This enables us to convert any basic modal formula into a first-order formula with one free variable. The standard translation is a bridge between the modal and classical worlds, a bridge that underlies much of the work of this chapter.

2.1 First steps in relational semantics

Suppose we have a set of proposition symbols (whose elements we typically write as p, q, r and so on) and a set of modality symbols (whose elements we typically write as m, m', m'', and so on). The choice of PROP and MOD is called the *signature* (or *similarity type*) of the language; in what follows we'll tacitly assume that PROP is denumerably infinite, and we'll often work with signatures in which MOD contains only a single element. Given a signature, we define the *basic modal language* (over the signature) as follows:

$$\varphi \quad ::= \quad p \mid \top \mid \bot \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \mid \varphi \leftrightarrow \psi \mid \langle m \rangle \varphi \mid [m]\varphi.$$

That is, a basic modal formula is either a proposition symbol, a boolean constant, a boolean combination of basic modal formulas, or (most interesting of all) a formula prefixed by a diamond

Example: Modal Logic Textbook

2 BASIC MODAL LOGIC

In this section we introduce the basic modal language and its relational semantics. We define basic modal syntax, introduce models and frames, and give the satisfaction definition. We then draw the reader's attention to the internal perspective that modal languages offer on relational structure, and explain why models and frames should be thought of as graphs. Following this we give the standard translation. This enables us to convert any basic modal formula into a first-order formula with one free variable. The standard translation is a bridge between the modal and classical worlds, a bridge that underlies much of the work of this chapter.

2.1 *First steps in relational semantics*

Syntax

Metalanguage

...ose elements we typically write as p, q, r and ...ents we typically write as m, m', m'', and so ...*nature* (*or similarity type*) of the language; in what follows we'll tacitly assume that PROP is denumerably infinite, and we'll often work with signatures in which MOD contains only a single element. Given a signature, we define the *basic modal language* (over the signature) as follows:

$$\varphi \quad ::= \quad p \mid \top \mid \bot \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \to \psi \mid \varphi \leftrightarrow \psi \mid \langle m \rangle \varphi \mid [m]\varphi.$$

That is, a basic modal formula is either a proposition symbol, a boolean constant, a boolean combination of basic modal formulas, or (most interesting of all) a formula prefixed by a diamond

Example: Modal Logic Textbook

A *model* (or *Kripke model*) $\mathfrak{M}$ for the basic modal language (over some fixed signature) is a triple $\mathfrak{M} = (W, \{R^m\}_{m \in \text{MOD}}, V)$. Here W, the *domain*, is a non-empty set, whose elements we usually call *points*, but which, for reasons which will soon be clear, are sometimes called *states*, *times*, *situations*, *worlds* and other things besides. Each R^m in a model is a binary relation on W, and V is a function (the valuation) that assigns to each proposition symbol p in PROP a subset $V(p)$ of W; think of $V(p)$ as the set of points in $\mathfrak{M}$ where p is true. The first two components $(W, \{R^m\}_{m \in \text{MOD}})$ of $\mathfrak{M}$ are called the *frame* underlying the model. If there is only one relation in the model, we typically write (W, R) for its frame, and (W, R, V) for the model itself. We encourage the reader to think of Kripke models as graphs (or to be slightly more precise, *directed graphs*, that is, graphs whose points are linked by directed arrows) and will shortly give some examples which show why this is helpful.

Suppose w is a point in a model $\mathfrak{M} = (W, \{R^m\}_{m \in \text{MOD}}, V)$. Then we inductively define the notion of a formula φ being *satisfied* (or *true*) in $\mathfrak{M}$ at point w as follows (we omit some of the clauses for the booleans):

$$
\begin{array}{rcl}
\mathfrak{M}, w \models p & \text{iff} & w \in V(p), \\
\mathfrak{M}, w \models \top & & \text{always}, \\
\mathfrak{M}, w \models \bot & & \text{never}, \\
\mathfrak{M}, w \models \neg\varphi & \text{iff} & \text{not } \mathfrak{M}, w \models \varphi \text{ (notation: } \mathfrak{M}, w \not\models \varphi), \\
\mathfrak{M}, w \models \varphi \wedge \psi & \text{iff} & \mathfrak{M}, w \models \varphi \text{ and } \mathfrak{M}, w \models \psi, \\
\mathfrak{M}, w \models \varphi \rightarrow \psi & \text{iff} & \mathfrak{M}, w \not\models \varphi \text{ or } \mathfrak{M}, w \models \psi, \\
\mathfrak{M}, w \models \langle m \rangle\varphi & \text{iff} & \text{for some } v \in W \text{ such that } R^m wv \text{ we have } \mathfrak{M}, v \models \varphi, \\
\mathfrak{M}, w \models [m]\varphi & \text{iff} & \text{for all } v \in W \text{ such that } R^m wv \text{ we have } \mathfrak{M}, v \models \varphi.
\end{array}
$$

Example: Modal Logic Textbook

A *model* (or *Kripke model*) $\mathfrak{M}$ for the basic modal language (over some fixed signature) is a triple $\mathfrak{M} = (W, \{R^m\}_{m \in \text{MOD}}, V)$. Here W, the *domain*, is a non-empty set, whose elements we usually call *points*, but which, for reasons which will soon be clear, are sometimes called *states*, *times* … in a model is a binary relation on W, and … position symbol p in PROP a subset $V(p)$ … p is true. The first two components $(W, \{$ … model. If there is only one relation in the … (W, R, V) for the model itself. We encourage the reader to think of Kripke models as graphs (or to be slightly more precise, *directed graphs*, that is, graphs whose points are linked by directed arrows) and will shortly give some examples which show why this is helpful.

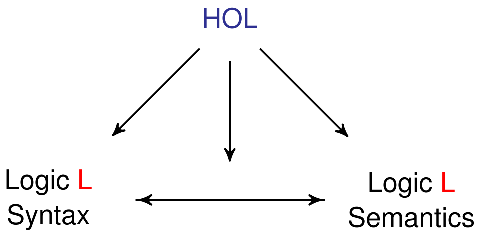
Metalanguage

Suppose w is a point in a model $\mathfrak{M} = (W, \{R^m\}_{m \in \text{MOD}}, V)$. Then we inductively define the notion of a formula φ being *satisfied* (or *true*) in $\mathfrak{M}$ at point w as follows (we omit some of the clauses for the booleans):

Semantics

$$
\begin{array}{rll}
\mathfrak{M}, w \models p & \text{iff} & w \in V(p), \\
\mathfrak{M}, w \models \top & & \text{always}, \\
\mathfrak{M}, w \models \bot & & \text{never}, \\
\mathfrak{M}, w \models \neg\varphi & \text{iff} & \text{not } \mathfrak{M}, w \models \varphi \text{ (notation: } \mathfrak{M}, w \not\models \varphi), \\
\mathfrak{M}, w \models \varphi \wedge \psi & \text{iff} & \mathfrak{M}, w \models \varphi \text{ and } \mathfrak{M}, w \models \psi, \\
\mathfrak{M}, w \models \varphi \to \psi & \text{iff} & \mathfrak{M}, w \not\models \varphi \text{ or } \mathfrak{M}, w \models \psi, \\
\mathfrak{M}, w \models \langle m \rangle \varphi & \text{iff} & \text{for some } v \in W \text{ such that } R^m wv \text{ we have } \mathfrak{M}, v \models \varphi, \\
\mathfrak{M}, w \models [m] \varphi & \text{iff} & \text{for all } v \in W \text{ such that } R^m wv \text{ we have } \mathfrak{M}, v \models \varphi.
\end{array}
$$

Universal Reasoning in Meta-Logic HOL



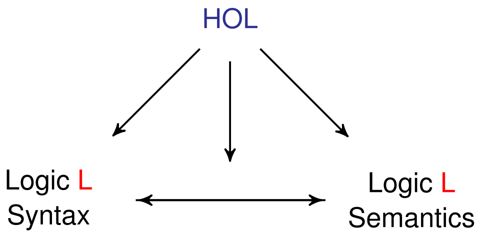
Examples for L we have already studied:
Intuitionistic Logics, (Mathematical) Fuzzy Logics, Free Logic, Modal Logics, Description Logics, Conditional Logics, Access Control Logics, Hybrid Logics, Multivalued Logics, Logics with Neighborhood Semantics, Paraconsistent Logics, Dyadic Deontic Logic, . . .

Embedding works also for quantifiers (first-order & higher-order)

HOL provers become universal logic reasoning engines!

interactive:	Isabelle/HOL, PVS, HOL4, Hol Light, Coq/HOL, . . .
automated:	Leo-III, LEO-II, Satallax, TPS, Nitpick, Isabelle/HOL, . . .

Universal Reasoning in Meta-Logic HOL



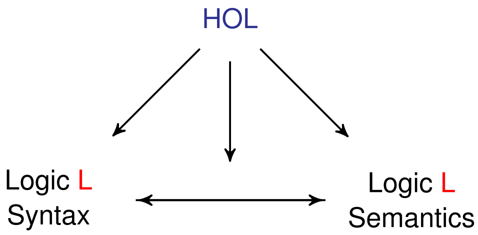
Examples for L we have already studied:
Intuitionistic Logics, (Mathematical) Fuzzy Logics, Free Logic, Modal Logics, Description Logics, Conditional Logics, Access Control Logics, Hybrid Logics, Multivalued Logics, Logics with Neighborhood Semantics, Paraconsistent Logics, Dyadic Deontic Logic, . . .

Embedding works also for quantifiers (first-order & higher-order)

HOL provers become universal logic reasoning engines!

interactive: Isabelle/HOL, PVS, HOL4, Hol Light, Coq/HOL, . . .

automated: Leo-III, LEO-II, Satallax, TPS, Nitpick, Isabelle/HOL, . . .



Examples for L we have already studied:
Intuitionistic Logics, (Mathematical) Fuzzy Logics, Free Logic, Modal Logics, Description Logics, Conditional Logics, Access Control Logics, Hybrid Logics, Multivalued Logics, Logics with Neighborhood Semantics, Paraconsistent Logics, Dyadic Deontic Logic, . . .

Embedding works also for quantifiers (first-order & higher-order)

HOL provers become universal logic reasoning engines!

interactive:	Isabelle/HOL, PVS, HOL4, Hol Light, Coq/HOL, . . .
automated:	Leo-III, LEO-II, Satallax, TPS, Nitpick, Isabelle/HOL, . . .

Isabelle/HOL (one of various Theorem Provers for HOL)

Isabelle

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

TUM TECHNISCHE UNIVERSITÄT MÜNCHEN

- Home
- Overview
- Installation
- Documentation

Site Mirrors:
Cambridge (.uk)
Munich (.de)
Sydney (.au)
Potsdam, NY (.us)

What is Isabelle?

Isabelle is a generic proof assistant. It allows mathematical formulas to be expressed in a formal language and provides tools for proving those formulas in a logical calculus. Isabelle was originally developed at the University of Cambridge and Technische Universität München, but now includes numerous contributions from institutions and individuals worldwide. See the Isabelle overview for a brief introduction.

Now available: Isabelle2017 (October 2017)

Download for Mac OS X

Download for Linux - Download for Windows (32bit) - Download for Windows (64bit) - Download for Mac OS X

Some notable changes:

- Experimental support for Visual Studio Code as alternative PIDE front-end.
- Improved Isabelle/jEdit Prover IDE: management of session sources independently of editor buffers, removal of unused theories, explicit indication of theory status, more careful auto-indentation.
- Session-qualified theory imports.
- Code generator improvements: support for statically embedded computations.
- Numerous HOL library improvements.
- More material in HOL-Algebra, HOL-Computational_Algebra and HOL-Analysis (ported from HOL-Light).
- Improved Nunchaku model finder, now in main HOL.
- SQL database support in Isabelle/Scala.

See also the cumulative NEWS.

Distribution & Support

Isabelle is distributed for free under a conglomerate of open-source licenses, but the main code-base is subject to BSD-style regulations. The application bundles include source and binary packages and documentation, see the detailed installation instructions. A vast collection of Isabelle examples and applications is available from the Archive of Formal Proofs.

`https://isabelle.in.tum.de`

many other systems:

Coq, HOL, HOL Light, PVS, Lean, NuPrL, IMPS, ACL2, **Leo-II/Leo-III**, . . .

Universal Reasoning in Isabelle/HOL

GodProof.thy

GodProof.thy (~/chris/trunk/tex/talks/2018-DEON/DEMO/)

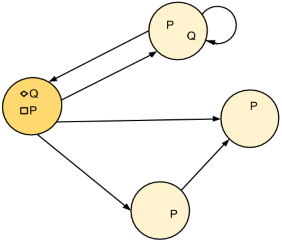
```
theory GodProof imports Main
begin
 typedecl i -- "type for possible worlds"
 typedecl μ -- "type for individuals"
 type_synonym σ = "(i⇒bool)"

(* Shallow embedding modal logic connectives in HOL *)
 abbreviation mneg ("¬_"[52]53)    where "¬φ    ≡ λw. ¬φ(w)"
 abbreviation mand (infixr"∧"51)   where "φ∧ψ  ≡ λw. φ(w)∧ψ(w)"
 abbreviation mor (infixr"∨"50)    where "φ∨ψ  ≡ λw. φ(w)∨ψ(w)"
 abbreviation mimp (infixr"→"49)   where "φ→ψ  ≡ λw. φ(w)⟶ψ(w)"
 abbreviation mequ (infixr"↔"48)   where "φ↔ψ  ≡ λw. φ(w)⟷ψ(w)"
 abbreviation mnegpred ("¬_"[52]53) where "¬Φ  ≡ λx.λw. ¬Φ(x)(w)"

(* Generic box and diamond operators *)
 abbreviation mboxgen ("□") where "□r φ ≡ λw. ∀v.  r w v ⟶ φ(v)"
 abbreviation mdiagen ("◇") where "◇r φ ≡ λw. ∃v. r w v ∧ φ(v)"

(* Shallow embedding of constant domain quantifiers in HOL *)
 abbreviation mall_const ("∀c") where "∀c Φ ≡ λw.∀x. Φ(x)(w)"
 abbreviation mallB_const (binder"∀c"[8]9) where "∀c x. φ(x) ≡ ∀c φ"
 abbreviation mexi_const ("∃c") where "∃c Φ ≡ λw.∃x. Φ(x)(w)"
 abbreviation mexiB_const (binder"∃c"[8]9) where "∃c x. φ(x) ≡ ∃c φ"

(* Global validity: truth in all possible worlds *)
 abbreviation mvalid :: "σ ⇒ bool" ("⌊_⌋"[7]110) where "⌊p⌋ ≡ ∀w. p w"

(* Shallow embedding of varying domain quantifiers in HOL *)
```

Documentation Sidekick State Theories

Output Query Sledgehammer Symbols

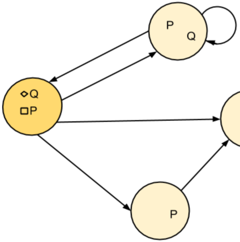
Universal Logic Reasoning in Isabelle/HOL

Properties of □ and ◇ correlated to structure of transition system between worlds



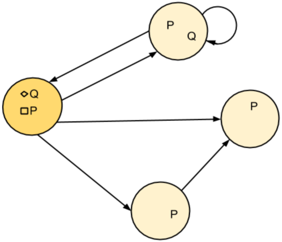
Universal Logic Reasoning in Isabelle/HOL

Properties of $\Box$ and $\Diamond$ correlated to structure of transition system between worlds



- Logic K: — (no restrictions, any structure)
- Logic M: reflexiv transition relation, $\forall P. \Box P \to P$
- Logic KB: symmetric transition relation, $\forall P. P \to \Box\Diamond P$
- Logic S5: equivelance relation as transition system, add $\forall P. \Box P \to \Box\Box P$

Properties of $\Box$ and $\Diamond$ correlated to structure of transition system between worlds



- Logic K: — (no restrictions, any structure)
- Logic M: reflexiv transition relation, $\forall P.\Box P \rightarrow P$
- Logic KB: symmetric transition relation, $\forall P.P \rightarrow \Box\Diamond P$
- Logic S5: equivelance relation as transition system, add $\forall P.\Box P \rightarrow \Box\Box P$

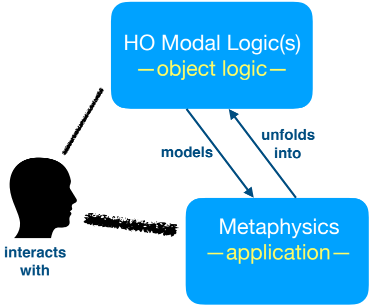
- Logic D: serial transition relation, $\forall P.\Box P \rightarrow \Diamond P$ (**S**tandard **D**eontic **L**ogic)

Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]

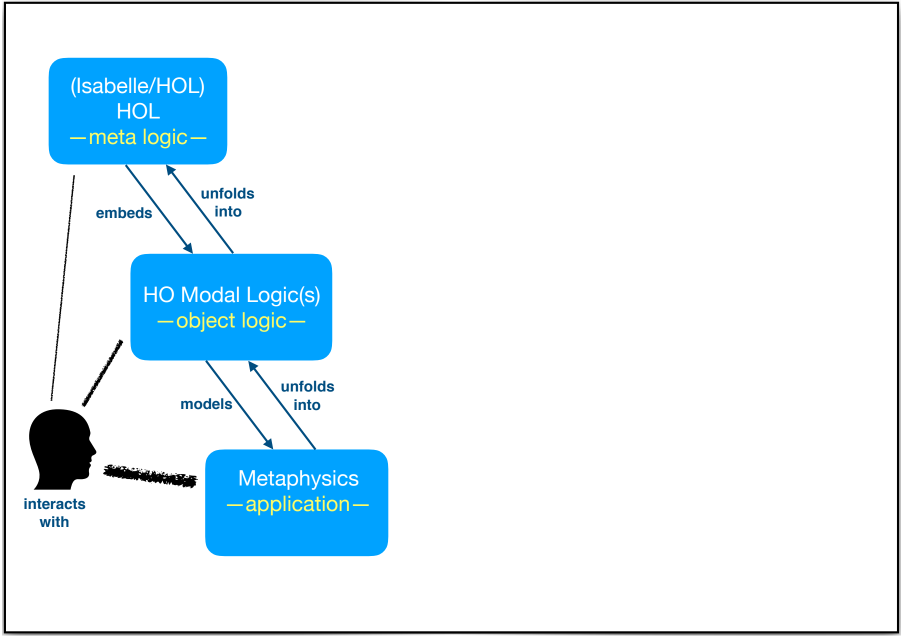
Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]



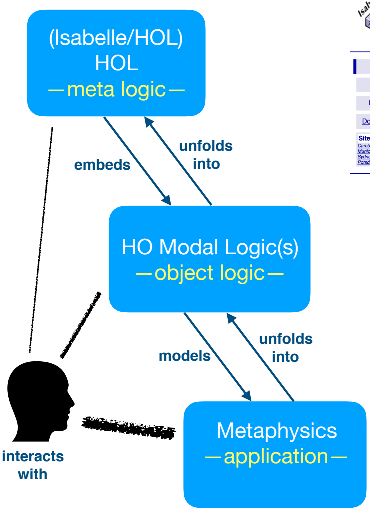
Universal (Meta-)Logical Reasoning Framework

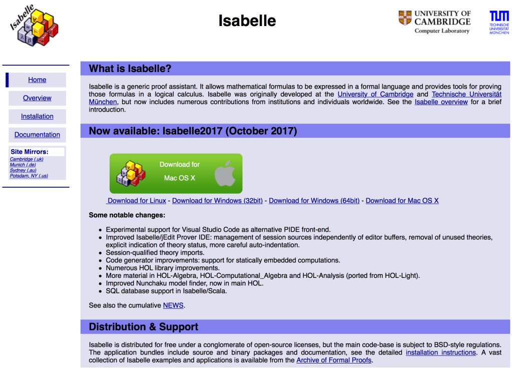
[Benzmüller_SBMF 2017]



Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]





Isabelle

UNIVERSITY OF CAMBRIDGE Computer Laboratory

Home

Overview

Installation

Documentation

Site Mirrors:
Cambridge (.uk)
Munich (.de)
Sydney (.au)
Potsdam, NY (.us)

What is Isabelle?

Isabelle is a generic proof assistant. It allows mathematical formulas to be expressed in a formal language and provides tools for proving those formulas in a logical calculus. Isabelle was originally developed at the University of Cambridge and Technische Universität München, but now includes numerous contributions from institutions and individuals worldwide. See the Isabelle overview for a brief introduction.

Now available: Isabelle2017 (October 2017)

Download for Mac OS X

Download for Linux - Download for Windows (32bit) - Download for Windows (64bit) - Download for Mac OS X

Some notable changes:

- Experimental support for Visual Studio Code as alternative PIDE front-end.
- Improved Isabelle/jEdit Prover IDE: management of session sources independently of editor buffers, removal of unused theories, explicit indication of theory status, more careful auto-indentation.
- Session-qualified theory imports.
- Code generator improvements: support for statically embedded computations.
- Numerous HOL library improvements.
- More material in HOL-Algebra, HOL-Computational_Algebra and HOL-Analysis (ported from HOL-Light).
- Improved Nunchaku model finder, now in main HOL.
- SQL database support in Isabelle/Scala.

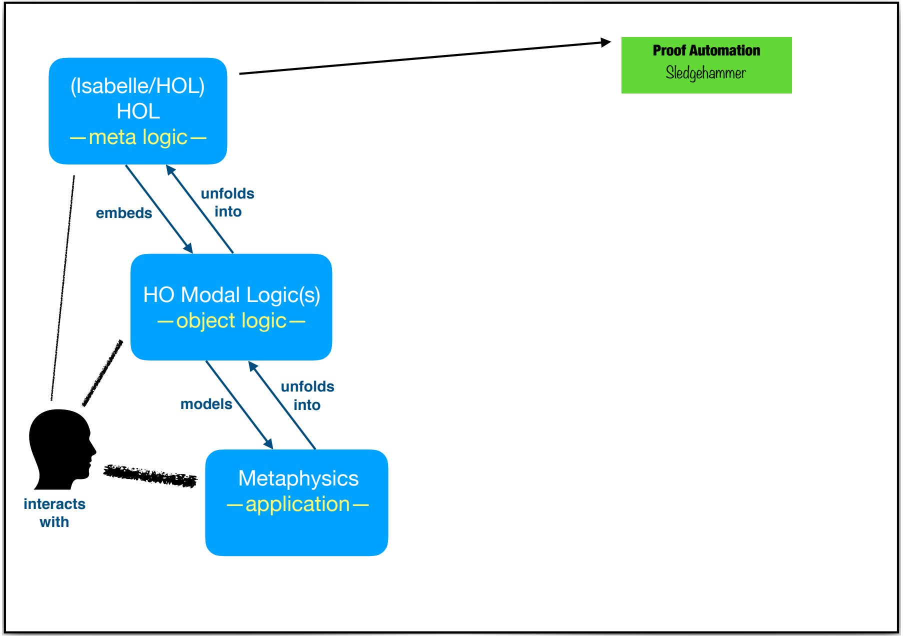
See also the cumulative NEWS.

Distribution & Support

Isabelle is distributed for free under a conglomerate of open-source licenses, but the main code-base is subject to BSD-style regulations. The application bundles include source and binary packages and documentation, see the detailed installation instructions. A vast collection of Isabelle examples and applications is available from the Archive of Formal Proofs.

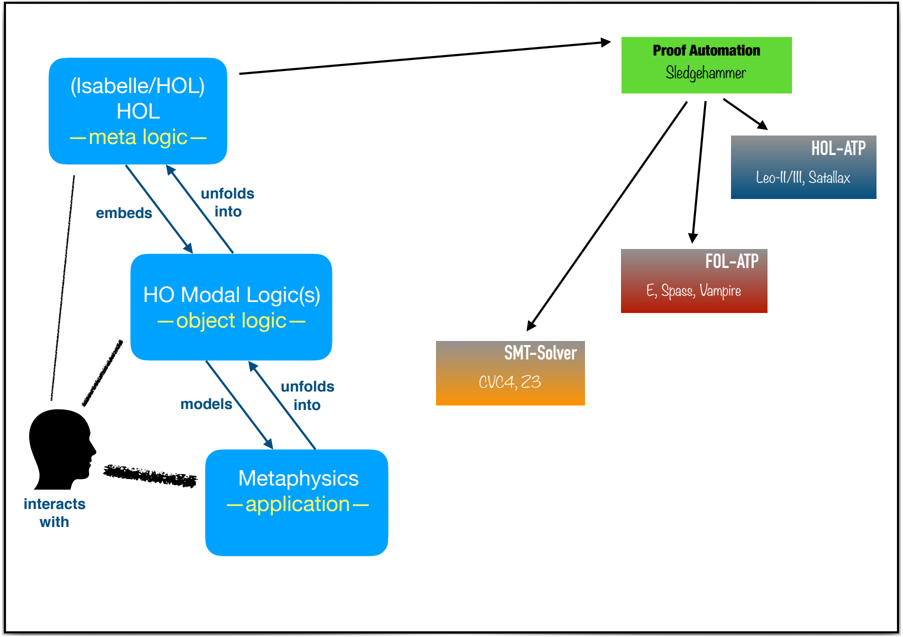
Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]



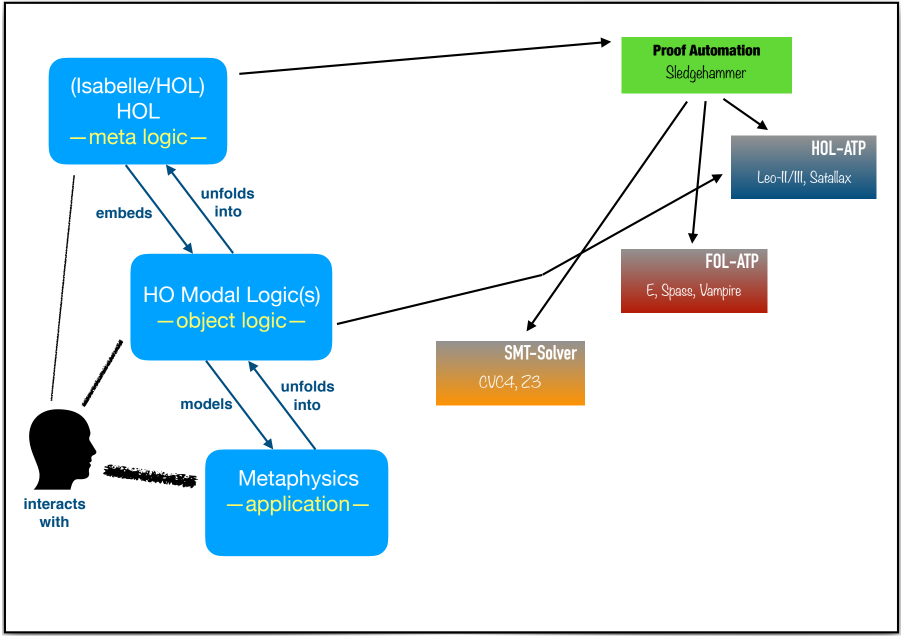
Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]



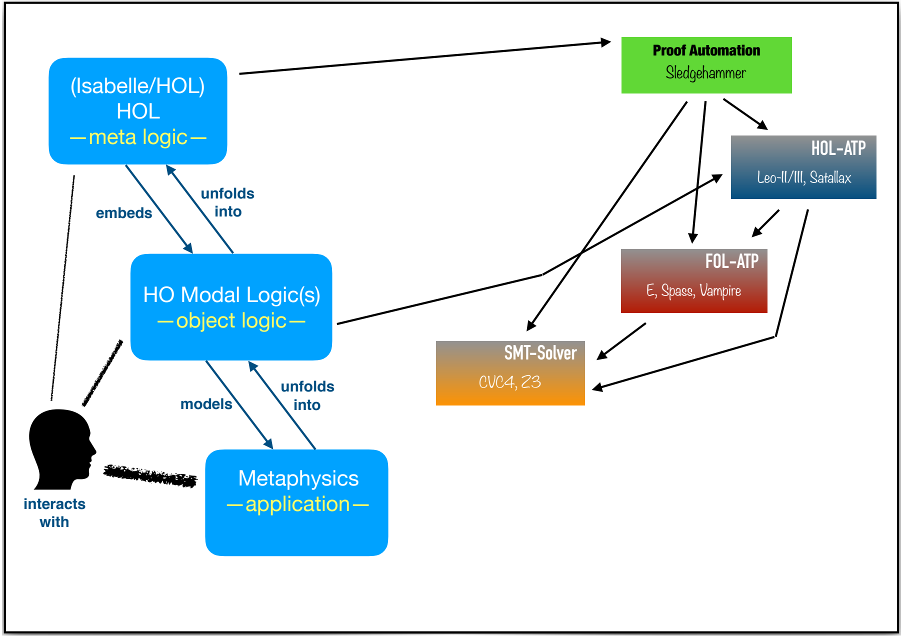
Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]



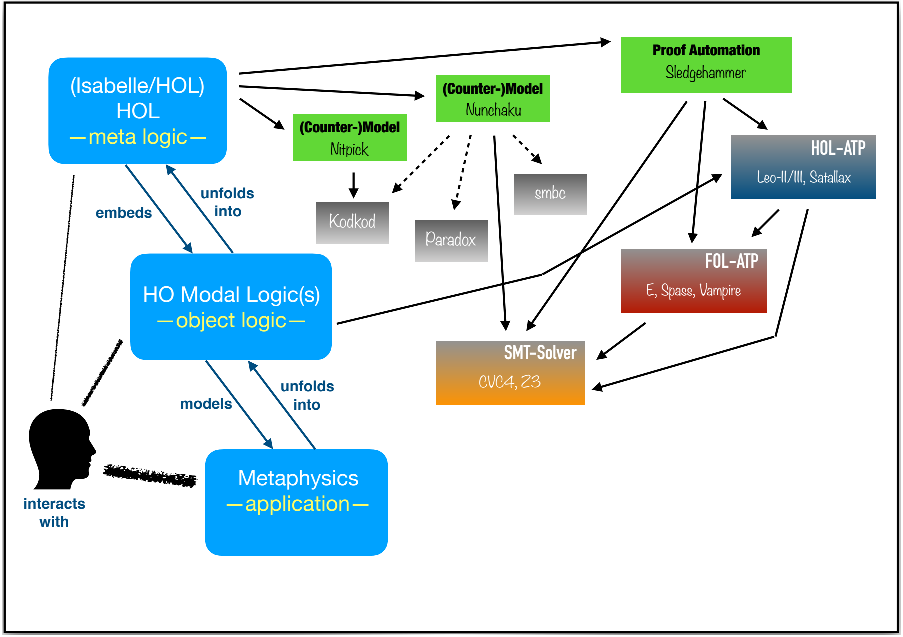
Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]

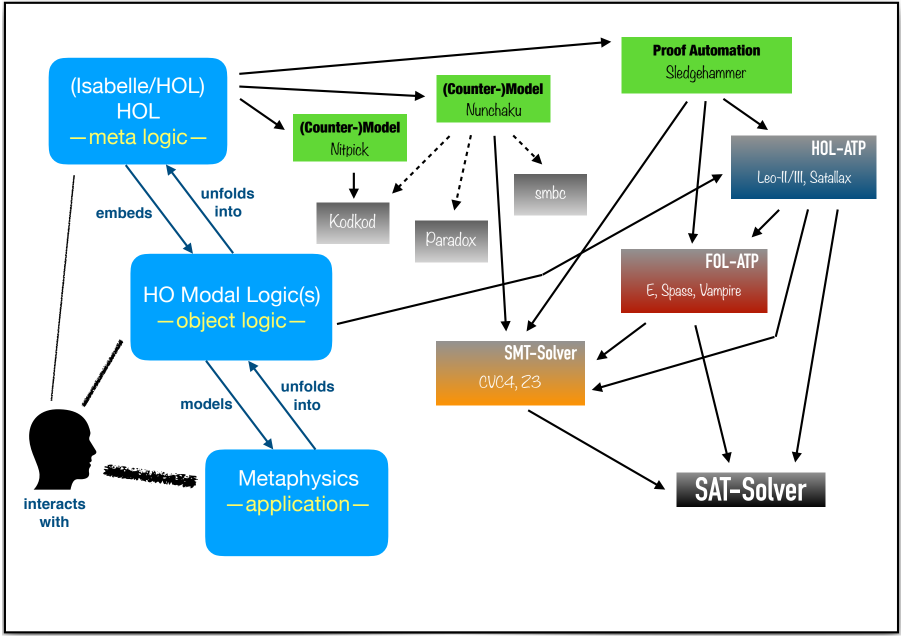


Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]

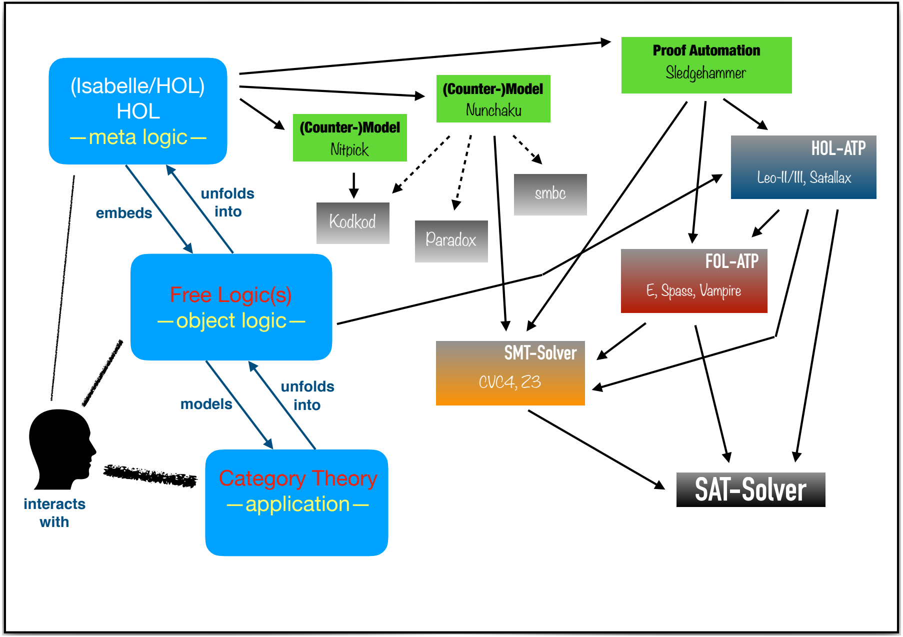


Universal (Meta-)Logical Reasoning Framework [Benzmüller_SBMF 2017]



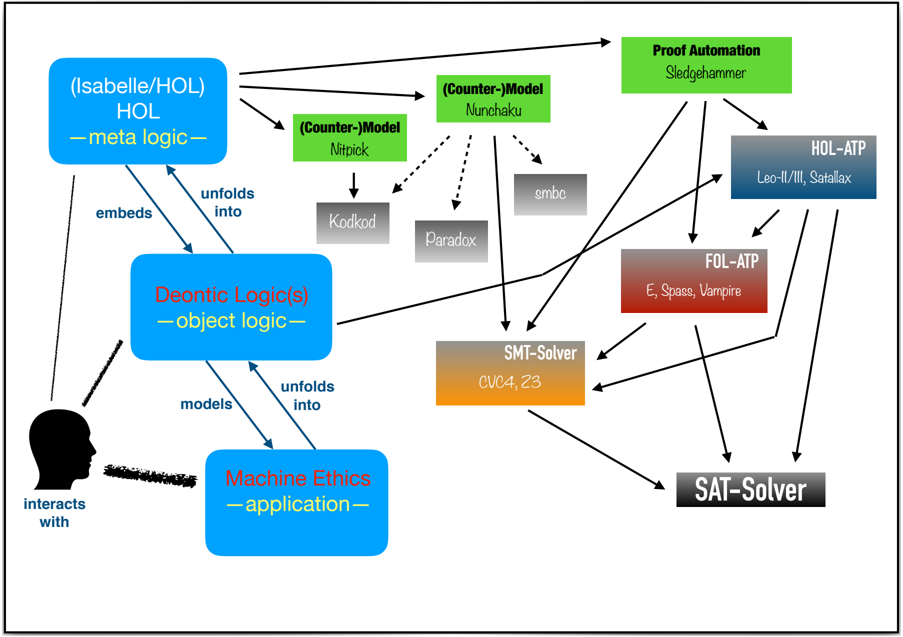
Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]



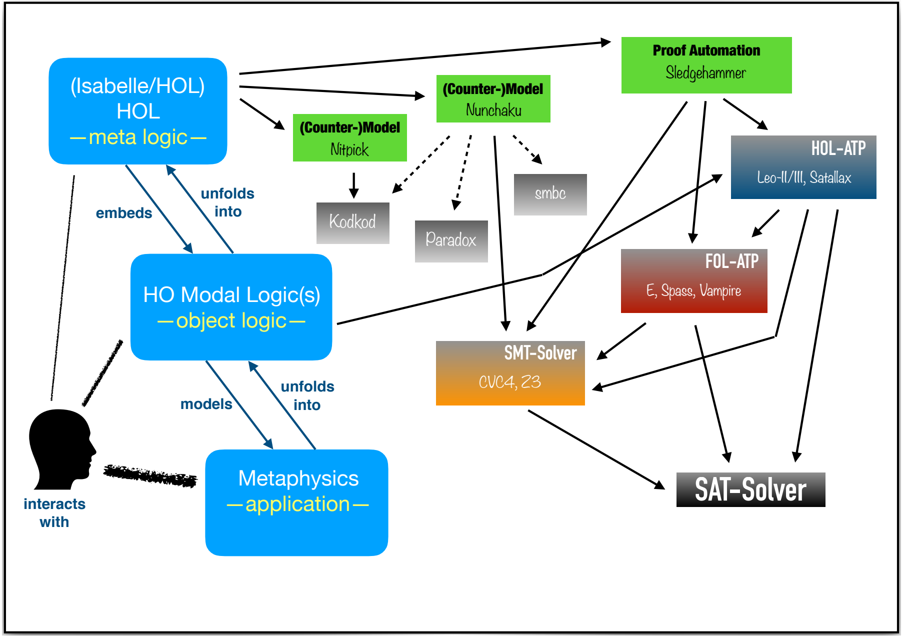
Universal (Meta-)Logical Reasoning Framework

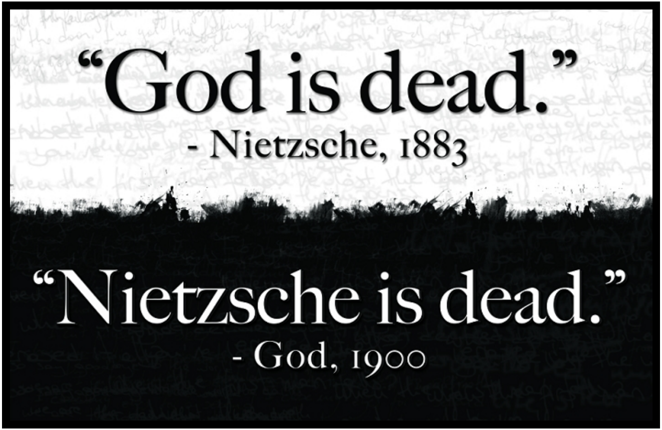
[Benzmüller_SBMF 2017]



Universal (Meta-)Logical Reasoning Framework

[Benzmüller_SBMF 2017]





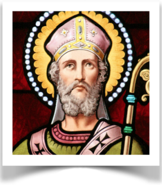
Part C

Evidence: Analysing Rational Arguments in Metaphysics

[BenzmüllerWoltzenlogelPaleo, ECAI, 2014 + IJCAI, 2016 + KI 2016 + . . .]

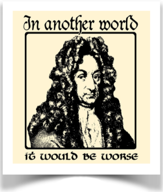
Ontological Proofs of God's Existence

A Long and Continuing Tradition in Philosophy



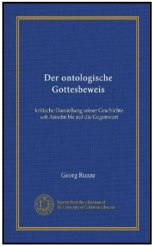
St. Anselm

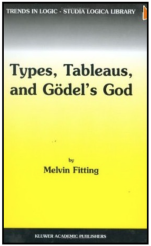
Descartes

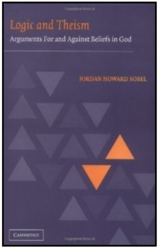


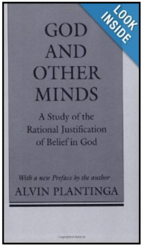
Leibniz

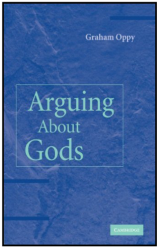
Gödel











Computational Metaphysics: Kurt Gödel's Ontological Argument

Ontologischer Beweis   Feb 10, 1970

$P(\varphi)$ φ is positive   (or $\varphi \in P$)

At. 1 $P(\varphi).P(\psi) \supset P(\varphi.\psi)$ *   At 2 $P(\varphi) \vee P(\sim\varphi)$

P1 $G(x) \equiv (\varphi)[P(\varphi) \supset \varphi(x)]$ (God)

P2 $\varphi \text{ Ess. } x \equiv (\psi)[\psi(x) \supset N(y)[\varphi(y) \supset \psi(y)]]$ (Essence of x)

$p \supset_N q = N(p \supset q)$ Necessity

At 2 $P(\varphi) \supset NP(\varphi)$, $\sim P(\varphi) \supset N \sim P(\varphi)$ } because it follows from the nature of the property

Th. $G(x) \supset G \text{ Ess. } x$

Df. $E(x) \equiv (\varphi)[\varphi \text{ Ess } x \supset N \exists x\, \varphi(x)]$ necessary Existence

Ax 3 $P(E)$

Th. $G(x) \supset N(\exists y)\, G(y)$

hence $(\exists x) G(x) \supset N(\exists y) G(y)$

" $M(\exists x) G(x) \supset MN(\exists y) G(y)$ M = possibility

" $\supset N(\exists y) G(y)$

any two essences of x are nec. equivalent

exclusive or * and for any number of summands

$M(\exists x) G(x)$ means all [the system] ... patible This is ...

At 4: $P(\varphi).\varphi \supset_N \psi$ ...

{ $x = x$ is p... ; $x \neq x$ is ...

But if a system S of ...

It would mean that the sum prop. s (which is positive) would be $x \neq x$

Positive means positive in the moral aesth. sense (independently of the accidental structure of the world). Only then the ax. true. It m... also means pure "attribution"* as opposed to "privation" (or containing privation). This interpret. simpler proof

If φ pos. not: $(x) N \sim \varphi(x)$. Otherwise: $\varphi(x) \supset_N x \neq x$ hence $x \neq x$ positive not $x = x$ ... contrary Ax ... or the exist. of pos. prop.

* i.e. the normal form in terms of elem. prop. contains a member without negation.

Computational Metaphysics: Kurt Gödel's Ontological Argument

Ontologischer Beweis — Feb 10, 1970

$P(\varphi)$ — φ is positive ($\varphi \in P$)

At. 1 $P(\varphi).P(\psi) \supset P(\varphi.\psi)$ • At 2 $P(\varphi) \vee P(\sim\varphi)$

P1 $G(x) \equiv (\varphi)[P(\varphi) \supset \varphi(x)]$ (God)

P2 $\varphi\ \mathrm{Ess.}\ x \equiv (\psi)[\psi(x) \supset N(y)[\varphi(y) \supset \psi(y)]]$ (Essence of x

$p \supset_N q = N(p \supset q)$ Necessity

At 2 $P(\varphi) \supset NP(\varphi)$, $\sim P(\varphi) \supset N \sim P(\varphi)$ } because it follows from the nature of the property

Th. $G(x) \supset G\ \mathrm{Ess.}\ x$

Df. $E(x) \equiv (\varphi)[\varphi\ \mathrm{Ess}\ x \supset N\ \exists x\ \varphi(x)]$ necessary Existence

Ax 3 $P(E)$

Th. $G(x) \supset N(\exists y)G(y)$

hence $(\exists x)G(x) \supset N(\exists y)G(y)$

" $M(\exists x)G(x) \supset MN(\exists y)G(y)$ — M = possibility

" $\supset N(\exists y)G(y)$

any two essences of x are nec. equivalent

exclusive or • and for any number of summands

$M(\exists x)G(x)$ means all pos. prop. is compatible [the system of] — This is true because of:

At 4: $P(\varphi).\varphi \supset_N \psi : \supset P(\psi)$ which impl

~~[illegible]~~ { $x = x$ is positive; $x \neq x$ is negative

But if a system S of pos. prop. were incom[patible] it would mean, that the sum prop. s (which is positive) would be $x \neq x$

Positive means positive in the moral aesth[etic] sense (independently of the accidental structure of the world). Only then the ax. true. It m[ay] also mean pure "attribution"[x] as opposed to "privation" (or containing privation). This interpret. simpler proof

If φ pos. not: $(x)N\sim\varphi(x)$. Otherwise: $\varphi(x) \supset_N x \neq x$ hence $x \neq x$ positive not $x = x$ neg. contrary At[...] or the exist. of pos. prop.

x i.e. the disj. normal form in terms of elem. prop. contains a member without negation.

Computational Metaphysics: Dana Scott's Variant

Axiom A1 Either a property or its negation is positive, but not both: $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2 A property necessarily implied by a positive property is positive: $\forall\phi\forall\psi[(P(\phi) \wedge \Box\forall x[\phi(x) \rightarrow \psi(x)]) \rightarrow P(\psi)]$

Thm. T1 Positive properties are possibly exemplified: $\forall\phi[P(\phi) \rightarrow \Diamond\exists x\phi(x)]$

Def. D1 A *God-like* being possesses all positive properties: $G(x) \leftrightarrow \forall\phi[P(\phi) \rightarrow \phi(x)]$

Axiom A3 The property of being God-like is positive: $P(G)$

Cor. C Possibly, God exists: $\Diamond\exists x G(x)$

Axiom A4 Positive properties are necessarily positive: $\forall\phi[P(\phi) \rightarrow \Box P(\phi)]$

Def. D2 An *essence* of an individual is a property possessed by it and necessarily implying any of its properties: $\phi\ ess.\ x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being: $\forall x[G(x) \rightarrow G\ ess.\ x]$

Def. D3 *Necessary existence* of an individual is the necessary exemplification of all its essences: $NE(x) \leftrightarrow \forall\phi[\phi\ ess.\ x \rightarrow \Box\exists y\phi(y)]$

Axiom A5 Necessary existence is a positive property: $P(NE)$

Thm. T3 Necessarily, God exists: $\Box\exists x G(x)$

Computational Metaphysics: Scott's Variant

Axiom A1 Either a property or its negation is positive, but not both: $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2 A property necessarily implied by a positive property is positive:
$$\forall\phi\forall\psi[(P(\phi) \wedge \Box\forall x[\phi(x) \to \psi(x)]) \to P(\psi)]$$

Thm. T1 Positive properties are possibly exemplified: $\forall\phi[P(\phi) \to \Diamond\exists x\phi(x)]$

Def. D1 A *God-like* being possesses all positive properties: $G(x) \leftrightarrow \forall\phi[P(\phi) \to \phi(x)]$

Axiom A3 The property of being God-like is positive: $P(G)$

Cor. C Possibly, God exists: $\Diamond\exists xG(x)$

Axiom A4 Positive properties are necessarily positive: $\forall\phi[P(\phi) \to \Box P(\phi)]$

Def. D2 An *essence* of an individual is a property possessed by it and necessarily implying any of its properties: $\phi\ ess.\ x \leftrightarrow \boxed{\phi(x) \wedge} \forall\psi(\psi(x) \to \Box\forall y(\phi(y) \to \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being: $\forall x[G(x) \to G\ ess.\ x]$

Def. D3 *Necessary existence* of an individual is the necessary exemplification of all its essences: $NE(x) \leftrightarrow \forall\phi[\phi\ ess.\ x \to \Box\exists y\phi(y)]$

Axiom A5 Necessary existence is a positive property: $P(NE)$

Thm. T3 Necessarily, God exists: $\Box\exists xG(x)$

Difference to Gödel (who omits this conjunct)

Computational Metaphysics: Scott's Variant of Gödel's Ontological Argument

Axiom A1 Either a property or its negation is positive, but not both: $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2 A property necessarily implied by a positive property is positive:
$$\forall\phi\forall\psi[(P(\phi) \wedge \boxed{\Box}\forall x[\phi(x) \rightarrow \psi(x)]) \rightarrow P(\psi)]$$

Thm. T1 Positive properties are possibly exemplified: $\forall\phi[P(\phi) \rightarrow \Diamond\exists x\phi(x)]$

Def. D1 A *God-like* being possesses all positive properties: $G(x) \leftrightarrow \forall\phi[P(\phi) \rightarrow \phi(x)]$

Axiom A3 The property of being God-like is positive: $P(G)$

Cor. C Possibly, God exists: $\boxed{\Diamond}\exists xG(x)$

Axiom A4 Positive properties are necessarily positive: $\forall\phi[P(\phi) \rightarrow \Box P(\phi)]$

Def. D2 An *essence* of an individual is a property possessed by it and necessarily implying any of its properties: $\phi\ ess.\ x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being: $\forall x[G(x) \rightarrow G\ ess.\ x]$

Def. D3 *Necessary existence* of an individual is the necessary exemplification of all its essences: $NE(x) \leftrightarrow \forall\phi[\phi\ ess.\ x \rightarrow \Box\exists y\phi(y)]$

Axiom A5 Necessary existence is a positive property: $P(NE)$

Thm. T3 Necessarily, God exists: $\Box\exists xG(x)$

Modal operators are used

Computational Metaphysics: Scott's Variant of Gödel's Ontological Argument

Axiom A1 Either a property or its negation is positive, but not both: $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2 A property necessarily implied by a positive property is positive: $\boxed{\forall\phi\forall\psi}[(P(\phi) \wedge \Box\forall x[\phi(x) \rightarrow \psi(x)]) \rightarrow P(\psi)]$

Thm. T1 Positive properties are possibly exemplified: $\forall\phi[P(\phi) \rightarrow \Diamond\exists x\phi(x)]$

Def. D1 A *God-like* being possesses all positive properties: $G(x) \leftrightarrow \boxed{\forall\phi}[P(\phi) \rightarrow \phi(x)]$

Axiom A3 The property of being God-like is positive: $P(G)$

Cor. C Possibly, God exists: $\Diamond\exists x G(x)$

Axiom A4 Positive properties are necessarily positive: $\forall\phi[P(\phi) \rightarrow \Box P(\phi)]$

Def. D2 An *essence* of an individual is a property possessed by it and necessarily implying any of its properties: $\phi\ ess.\ x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being: $\forall x[G(x) \rightarrow G\ ess.\ x]$

Def. D3 *Necessary existence* of an individual is the necessary exemplification of all its essences: $NE(x) \leftrightarrow \forall\phi[\phi\ ess.\ x \rightarrow \Box\exists y\phi(y)]$

Axiom A5 Necessary existence is a positive property: $P(NE)$

Thm. T3 Necessarily, God exists: $\Box\exists x G(x)$

second-order quantifiers

Computational Metaphysics: Scott's and Gödel's Variants — Demo

Axiom A1 Either a property or its negation is positive, but not both: $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2 A property necessarily implied by a positive property is positive:
$$\forall\phi\forall\psi[(P(\phi) \wedge \Box\forall x[\phi(x) \rightarrow \psi(x)]) \rightarrow P(\psi)]$$

Thm. T1 Positive properties are possibly exemplified: $\forall\phi[P(\phi) \rightarrow \Diamond\exists x\phi(x)]$

Def. D1 A *God-like* being possesses all positive properties: $G(x) \leftrightarrow \forall\phi[P(\phi) \rightarrow \phi(x)]$

Axiom A3 The property of being God-like is positive: $P(G)$

Cor. C Possibly, God exists: $\Diamond\exists x G(x)$

Axiom A4 Positive properties are necessarily positive: $\forall\phi[P(\phi) \rightarrow \Box P(\phi)]$

Def. D2 An *essence* of an individual is a property possessed by it and necessarily implying any of its properties: $\phi\ ess.\ x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being: $\forall x[G(x) \rightarrow G\ ess.\ x]$

Def. D3 *Necessary existence* of an individual is the necessary exemplification of all its essences: $NE(x) \leftrightarrow \forall\phi[\phi\ ess.\ x \rightarrow \Box\exists y\phi(y)]$

Axiom A5 Necessary existence is a positive property: $P(NE)$

Thm. T3 Necessarily, God exists: $\Box\exists x G(x)$

Computational Metaphysics: Scott's and Gödel's Variants — Demo

Axiom A1 $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2

$$\forall\phi\forall\psi[(P(\phi) \wedge \Box\forall x[\phi(x) \rightarrow \psi(x)]) \rightarrow P(\psi)]$$

Thm. T1 $\forall\phi[P(\phi) \rightarrow \Diamond\exists x\phi(x)]$

Def. D1 $G(x) \leftrightarrow \forall\phi[P(\phi) \rightarrow \phi(x)]$

Axiom A3 $P(G)$

Cor. C $\Diamond\exists x G(x)$

Axiom A4 $\forall\phi[P(\phi) \rightarrow \Box P(\phi)]$

Def. D2

$$\phi\ ess.\ x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$$

Thm. T2 $\forall x[G(x) \rightarrow G\ ess.\ x]$

Def. D3

$$NE(x) \leftrightarrow \forall\phi[\phi\ ess.\ x \rightarrow \Box\exists y\phi(y)]$$

Axiom A5 $P(NE)$

Thm. T3 $\Box\exists x G(x)$

Computational Metaphysics: Scott's and Gödel's Variants — Demo

Axiom A1 $\forall\phi[P(\neg\phi) \leftrightarrow \neg P(\phi)]$

Axiom A2

$$\forall\phi\forall\psi[(P(\phi) \wedge \Box\forall x[\phi(x) \rightarrow \psi(x)]) \rightarrow P(\psi)]$$

Def. D1 $G(x) \leftrightarrow \forall\phi[P(\phi) \rightarrow \phi(x)]$

Axiom A3 $P(G)$

Axiom A4 $\forall\phi[P(\phi) \rightarrow \Box P(\phi)]$

Def. D2

$$\phi \; ess. \; x \leftrightarrow \phi(x) \; \wedge \; \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$$

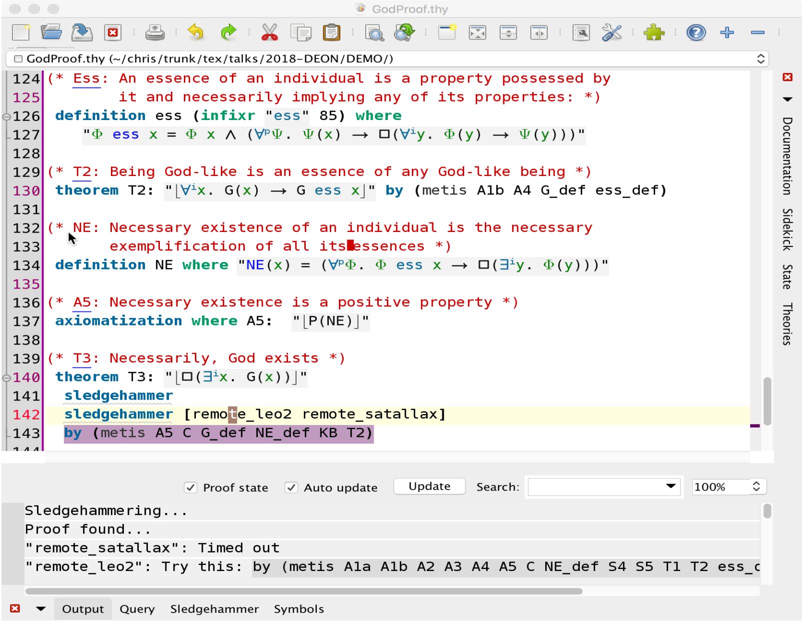
Def. D3

$$NE(x) \leftrightarrow \forall\phi[\phi \; ess. \; x \rightarrow \Box\exists y\phi(y)]$$

Axiom A5 $P(NE)$

Thm. T3 $\Box\exists x G(x)$

Computational Metaphysics: Scott's and Gödel's Variants — Demo



Results of our Experiments

Variant of Dana Scott

- the premises are **consistent**
- all argument steps are **logically correct** in (higher-order, extensional) modal logic
  - correct in logic **S5**
  - weaker logic **KB** is already sufficient
  - philosophical critique about use of S5 not justfied
- minimal dependencies determined by theorem provers
- alternative proofs (different from the ones in literature)

Intermediate Conclusion:

With our technology. . .
. . . it is possible to verify (selected) masterpiece arguments in philosophy

Results of our Experiments

Variant of Dana Scott

- the premises are **consistent**
- all argument steps are **logically correct**
  in (higher-order, extensional) modal logic
  - correct in logic **S5**
  - weaker logic **KB** is already sufficient
  - philosophical critique about use of S5 not justfied
- minimal dependencies determined by theorem provers
- alternative proofs (different from the ones in literature)

Intermediate Conclusion:

With our technology. . .
. . . it is possible to verify (selected) masterpiece arguments in philosophy

Results of our Experiments

Variant of Dana Scott

- the premises are **consistent**
- all argument steps are **logically correct** in (higher-order, extensional) modal logic
  - correct in logic **S5**
  - weaker logic **KB** is already sufficient
  - philosophical critique about use of S5 not justfied
- minimal dependencies determined by theorem provers
- alternative proofs (different from the ones in literature)

Intermediate Conclusion:

With our technology. . .
. . . it is possible to verify (selected) masterpiece arguments in philosophy

Results of our Experiments

Variant of Dana Scott

- the premises are **consistent**
- all argument steps are **logically correct** in (higher-order, extensional) modal logic
  - correct in logic **S5**
  - weaker logic **KB** is already sufficient
  - philosophical critique about use of S5 not justfied
- minimal dependencies determined by theorem provers
- alternative proofs (different from the ones in literature)

Intermediate Conclusion:

With our technology. . .

. . . it is possible to verify (selected) masterpiece arguments in philosophy

Variant of Dana Scott

- the premises are **consistent**
- all argument steps are **logically correct**
  in (higher-order, extensional) modal logic
  - correct in logic **S5**
  - weaker logic **KB** is already sufficient
  - philosophical critique about use of S5 not justfied
- minimal dependencies determined by theorem provers
- alternative proofs (different from the ones in literature)

Intermediate Conclusion:

With our technology. . .
. . . it is possible to verify (selected) masterpiece arguments in philosophy

Results of our Experiments

Variant of Kurt Gödel

- the premises are inconsistent/contradictory
- everything follows!
- Philosophers had not seen this
- . . . but my theorem prover LEO-II did

Intermediate Conclusion:

Our technology . . .
. . . can reveal flawed arguments and can even contribute new knowledge.

Results of our Experiments

Variant of Kurt Gödel

- the premises are inconsistent/contradictory
- everything follows!
- Philosophers had not seen this
- . . . but my theorem prover LEO-II did

Intermediate Conclusion:

Our technology . . .
. . . can reveal flawed arguments and can even contribute new knowledge.

Results of our Experiments

Variant of Kurt Gödel

- the premises are inconsistent/contradictory
- everything follows!
- Philosophers had not seen this
- . . . but my theorem prover LEO-II did

Intermediate Conclusion:

Our technology . . .
. . . can reveal flawed arguments and can even contribute new knowledge.

Results of our Experiments

Variant of Kurt Gödel

- the premises are inconsistent/contradictory
- everything follows!
- Philosophers had not seen this
- ... but my theorem prover LEO-II did

Intermediate Conclusion:

Our technology ...
... can reveal flawed arguments and can even contribute new knowledge.

. . . we continue with Scott's version

Further corollaries we can prove

- Monotheism
- Gott is flawless (has only positive properties)
- . . .
- Modal collapse: $\varphi \longrightarrow \Box \varphi$

  - there are no contingent truths
  - no alternative worlds
  - everything is determined
  - no free will

Challenge: Can the Modal Collapse be avoided (with minimal changes)?

Results of our Analysis

. . . we continue with Scott's version

Further corollaries we can prove

- Monotheism
- Gott is flawless (has only positive properties)
- . . .
- Modal collapse: $\varphi \longrightarrow \Box \varphi$
  - there are no contingent truths
  - no alternative worlds
  - everything is determined
  - no free will

Challenge: Can the Modal Collapse be avoided (with minimal changes)?

Results of our Analysis

. . . we continue with Scott's version

Further corollaries we can prove

- Monotheism
- Gott is flawless (has only positive properties)
- . . .
- Modal collapse: $\varphi \longrightarrow \Box \varphi$

  - there are no contingent truths
  - no alternative worlds
  - everything is determined
  - no free will

Challenge: Can the Modal Collapse be avoided (with minimal changes)?

Can the Modal Collapse be avoided?

SOME EMENDATIONS OF GÖDEL'S ONTOLOGICAL PROOF

C. Anthony Anderson

Kurt Gödel's version of the ontological argument was shown by J. Howard Sobel to be defective, but some plausible modifications in the argument result in a version which is immune to Sobel's objection. A definition is suggested which permits the proof of some of Gödel's axioms.

Gödel's Ontological Proof Revisited *

C. Anthony Anderson and Michael Gettings

University of California, Santa Barbara
Department of Philosophy

Gödel's version of the modal ontological argument for the existence of God has been criticized by J. Howard Sobel [5] and modified by C. Anthony Anderson [1]. In the present paper we consider the extent to which Anderson's emendation is defeated by the type of objection first offered by the Monk Gaunilo to St. Anselm's original Ontological Argument. And we try to push the analysis of this Gödelian argument a bit further to bring it into closer agreement with the details of Gödel's own formulation. Finally, we indicate what seems to be the main weakness of this emendation of Gödel's attempted proof.

Magari and others on Gödel's ontological proof

Petr Hájek
Institute of Computer Science, Academy of Sciences
182 07 Prague, Czech Republic
e-mail: hajek@uivt.cas.cz

1 Introduction

This paper is a continuation of my paper [H] and concentrates almost exclusively to mathematical properties of logical systems underlying Gödel's ontological proof [G] and its variant by Anderson [A], with special care paid to Magari's criticism [M]. Since [H] is written in German, we shall try to summarize its content in such a way that knowledge of [H] will be not obligatory for reading the present paper (even it remains advantageous). Here we describe

Der Mathematiker und die Frage der Existenz Gottes
(betreffend Gödels ontologischen Beweis)

Es ist gut, daß wir nicht wissen,
sondern glauben, daß ein Gott sei.
(Kant, Nachlaß)

1. Einführung

Gödels zu Lebzeiten unveröffentlichter Beweis für die notwendige Existenz eines Gott-ähnlichen Wesens hat sowohl philosophisches als auch mathematisches Interesse geweckt. Zweck der vorliegenden Arbeit ist es, zu einer Deutung des Gödelschen Textes beizutragen, 1. durch Kommentierung der einschlägigen Literatur und 2. durch Bereitstellung von etwas Modelltheorie. Die Arbeit enthält keinen philosophischen Beitrag. Während der letzten Jahre habe ich etliche Male über Gödels Gottesbeweis vorgetragen, insbesondere auf dem Symposium zur Feier von Professor Gert Müller (Heidelberg, Jänner 1991), doch habe ich niemals beabsichtigt, eine Veröffentlichung über das Thema zu machen. Da ich wiederholt um eine schriftliche Version gebeten wurde, entschloß ich mich, schnell eine „erweiterte Kurzfassung"[1] zu schreiben, ohne aus ihr einen

Petr Hájek

A New Small Emendation of Gödel's Ontological Proof

Keywords: Ontological proof, Gödel, modal logic, comprehension, positive properties.

1. Introduction

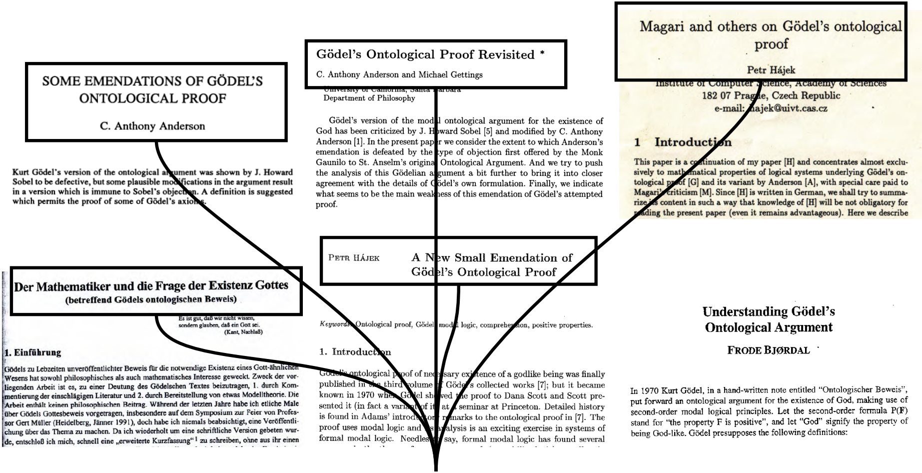
Gödel's ontological proof of necessary existence of a godlike being was finally published in the third volume of Gödel's collected works [7]; but it became known in 1970 when Gödel showed the proof to Dana Scott and Scott presented it (in fact a variant of it) at a seminar at Princeton. Detailed history is found in Adams' introductory remarks to the ontological proof in [7]. The proof uses modal logic and its analysis is an exciting exercise in systems of formal modal logic. Needles to say, formal modal logic has found several

Understanding Gödel's Ontological Argument

Frode Bjørdal

In 1970 Kurt Gödel, in a hand-written note entitled "Ontologischer Beweis", put forward an ontological argument for the existence of God, making use of second-order modal logical principles. Let the second-order formula P(F) stand for "the property F is positive", and let "God" signify the property of being God-like. Gödel presupposes the following definitions:

Can the Modal Collapse be avoided?



SOME EMENDATIONS OF GÖDEL'S ONTOLOGICAL PROOF

C. Anthony Anderson

Kurt Gödel's version of the ontological argument was shown by J. Howard Sobel to be defective, but some plausible modifications in the argument result in a version which is immune to Sobel's objection. A definition is suggested which permits the proof of some of Gödel's axioms.

Gödel's Ontological Proof Revisited *

C. Anthony Anderson and Michael Gettings

University of California, Santa Barbara
Department of Philosophy

Gödel's version of the modal ontological argument for the existence of God has been criticized by J. Howard Sobel [5] and modified by C. Anthony Anderson [1]. In the present paper we consider the extent to which Anderson's emendation is defeated by the type of objection first offered by the Monk Gaunilo to St. Anselm's original Ontological Argument. And we try to push the analysis of this Gödelian argument a bit further to bring it into closer agreement with the details of Gödel's own formulation. Finally, we indicate what seems to be the main weakness of this emendation of Gödel's attempted proof.

Magari and others on Gödel's ontological proof

Petr Hájek

Institute of Computer Science, Academy of Sciences
182 07 Prague, Czech Republic
e-mail: hajek@uivt.cas.cz

1 Introduction

This paper is a continuation of my paper [H] and concentrates almost exclusively to mathematical properties of logical systems underlying Gödel's ontological proof of [G] and its variant by Anderson [A], with special care paid to Magari's criticism [M]. Since [H] is written in German, we shall try to summarize its content in such a way that knowledge of [H] will be not obligatory for reading the present paper (even it remains advantageous). Here we describe

Der Mathematiker und die Frage der Existenz Gottes
(betreffend Gödels ontologischen Beweis)

Es ist gut, daß wir nicht wissen, sondern glauben, daß ein Gott sei.
(Kant, Nachlaß)

1. Einführung

Gödels zu Lebzeiten unveröffentlichter Beweis für die notwendige Existenz eines Gott-ähnlichen Wesens hat sowohl philosophisches als auch mathematisches Interesse geweckt. Zweck der vorliegenden Arbeit ist es, zu einer Deutung des Gödelschen Textes beizutragen, 1. durch Kommentierung der einschlägigen Literatur und 2. durch Bereitstellung von etwas Modelltheorie. Die Arbeit enthält keinen philosophischen Beitrag. Während der letzten Jahre habe ich etliche Male über Gödels Gottesbeweis vorgetragen, insbesondere auf dem Symposium zur Feier von Professor Gert Müller (Heidelberg, Jänner 1991), doch habe ich niemals beabsichtigt, eine Veröffentlichung über das Thema zu machen. Da ich wiederholt um eine schriftliche Version gebeten wurde, entschloß ich mich, schnell eine „erweiterte Kurzfassung"[1] zu schreiben, ohne aus ihr einen

PETR HÁJEK   A New Small Emendation of Gödel's Ontological Proof

Keywords: Ontological proof, Gödel, modal logic, comprehension, positive properties.

1. Introduction

Gödel's ontological proof of necessary existence of a godlike being was finally published in the third volume of Gödel's collected works [7]; but it became known in 1970 when Gödel showed the proof to Dana Scott and Scott presented it (in fact a variant of it) at a seminar at Princeton. Detailed history is found in Adams' introductory remarks to the ontological proof in [7]. The proof uses modal logic and its analysis is an exciting exercise in systems of formal modal logic. Needless to say, formal modal logic has found several

Understanding Gödel's Ontological Argument

FRODE BJØRDAL

In 1970 Kurt Gödel, in a hand-written note entitled "Ontologischer Beweis", put forward an ontological argument for the existence of God, making use of second-order modal logical principles. Let the second-order formula P(F) stand for "the property F is positive", and let "God" signify the property of being God-like. Gödel presupposes the following definitions:

— contributed to clarification of controversy —

— revealed various flaws and issues —

[Logica Universalis, 2017]

Comparison of

- Gödel/Scott (1972) — modal collapse
- C. Anthony Anderson (1990) — avoids modal collapse
- Melvin Fitting (2002) — avoids modal collapse

Questions:

- How do Anderson and Fitting the avoid modal collapse?
- Are their solutions related?

To answer this questions we will apply some notions from

- mathematics: ultrafilters
- philosophy of language: extension and intension of predicates

Summary of Results

- "Godlike" has been defined in terms of "positive properties"
- "positive properties" linked in experiments with notion of "ultrafilter"
- We then distinguished between
  - $\mathcal{P}$**:** positive intensional properties
  - $\mathcal{P}'$**:** positive ("rigidly intensionalised") extensions of properties
- Gödel/Scott variant axiomatises $\mathcal{P}$: $\mathcal{P} = \mathcal{P}'$ is an ultrafilter
- Anderson's variant axiomatises $\mathcal{P}$: $\mathcal{P} \neq \mathcal{P}'$; only $\mathcal{P}'$ is an ultrafilter
- Fitting's variant axiomatises only $\mathcal{P}'$: $\mathcal{P}'$ is an ultrafilter

> Modal collapse holds for Gödel/Scott variant, but not for Anderson's & Fitting's!
>
> They achieve this in seemingly different ways.
>
> Mathematically, however, their solutions appear closely related.

Summary of Results

- "Godlike" has been defined in terms of "positive properties"
- "positive properties" linked in experiments with notion of "ultrafilter"
- We then distinguished between

  $\mathcal{P}$: positive intensional properties
  $\mathcal{P}'$: positive ("rigidly intensionalised") extensions of properties
- Gödel/Scott variant axiomatises $\mathcal{P}$: $\mathcal{P} = \mathcal{P}'$ is an ultrafilter
- Anderson's variant axiomatises $\mathcal{P}$: $\mathcal{P} \neq \mathcal{P}'$; only $\mathcal{P}'$ is an ultrafilter
- Fitting's variant axiomatises only $\mathcal{P}'$: $\mathcal{P}'$ is an ultrafilter

Modal collapse holds for Gödel/Scott variant, but not for Anderson's & Fitting's!

They achieve this in seemingly different ways.

Mathematically, however, their solutions appear closely related.

Summary of Results

- "Godlike" has been defined in terms of "positive properties"
- "positive properties" linked in experiments with notion of "ultrafilter"
- We then distinguished between

  $\mathcal{P}$**:** positive intensional properties
  $\mathcal{P}'$**:** positive ("rigidly intensionalised") extensions of properties
- Gödel/Scott variant axiomatises $\mathcal{P}$: $\mathcal{P} = \mathcal{P}'$ is an ultrafilter
- Anderson's variant axiomatises $\mathcal{P}$: $\mathcal{P} \neq \mathcal{P}'$; only $\mathcal{P}'$ is an ultrafilter
- Fitting's variant axiomatises only $\mathcal{P}'$: $\mathcal{P}'$ is an ultrafilter

Modal collapse holds for Gödel/Scott variant, but not for Anderson's & Fitting's!

They achieve this in seemingly different ways.

Mathematically, however, their solutions appear closely related.

Summary of Results

- "Godlike" has been defined in terms of "positive properties"
- "positive properties" linked in experiments with notion of "ultrafilter"
- We then distinguished between

  $\mathcal{P}$: positive intensional properties

  $\mathcal{P}'$: positive ("rigidly intensionalised") extensions of properties
- Gödel/Scott variant axiomatises $\mathcal{P}$: $\mathcal{P} = \mathcal{P}'$ is an ultrafilter
- Anderson's variant axiomatises $\mathcal{P}$: $\mathcal{P} \neq \mathcal{P}'$; only $\mathcal{P}'$ is an ultrafilter
- Fitting's variant axiomatises only $\mathcal{P}'$: $\mathcal{P}'$ is an ultrafilter

Modal collapse holds for Gödel/Scott variant, but not for Anderson's & Fitting's!

They achieve this in seemingly different ways.

Mathematically, however, their solutions appear closely related.

Summary of Results

- "Godlike" has been defined in terms of "positive properties"
- "positive properties" linked in experiments with notion of "ultrafilter"
- We then distinguished between

  $\mathcal{P}$: positive intensional properties

  $\mathcal{P}'$: positive ("rigidly intensionalised") extensions of properties
- Gödel/Scott variant axiomatises $\mathcal{P}$: $\mathcal{P} = \mathcal{P}'$ is an ultrafilter
- Anderson's variant axiomatises $\mathcal{P}$: $\mathcal{P} \neq \mathcal{P}'$; only $\mathcal{P}'$ is an ultrafilter
- Fitting's variant axiomatises only $\mathcal{P}'$: $\mathcal{P}'$ is an ultrafilter

Modal collapse holds for Gödel/Scott variant, but not for Anderson's & Fitting's!

They achieve this in seemingly different ways.

Mathematically, however, their solutions appear closely related.

Summary of Results

- "Godlike" has been defined in terms of "positive properties"
- "positive properties" linked in experiments with notion of "ultrafilter"
- We then distinguished between

  $\mathcal{P}$: positive intensional properties

  $\mathcal{P}'$: positive ("rigidly intensionalised") extensions of properties
- Gödel/Scott variant axiomatises $\mathcal{P}$: $\mathcal{P} = \mathcal{P}'$ is an ultrafilter
- Anderson's variant axiomatises $\mathcal{P}$: $\mathcal{P} \neq \mathcal{P}'$; only $\mathcal{P}'$ is an ultrafilter
- Fitting's variant axiomatises only $\mathcal{P}'$: $\mathcal{P}'$ is an ultrafilter

Modal collapse holds for Gödel/Scott variant, but not for Anderson's & Fitting's!

They achieve this in seemingly different ways.

Mathematically, however, their solutions appear closely related.

Further Ongoing Experiments

Ed Zalta (Stanford)

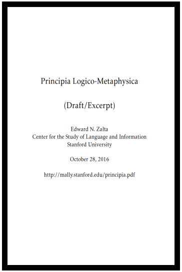
Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected

Daniel Kirchner
(Mathematics, FU Berlin)

Further Ongoing Experiments



Ed Zalta (Stanford)

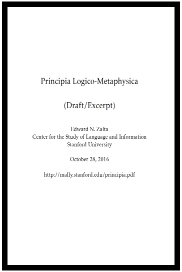
Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected

Daniel Kirchner
(Mathematics, FU Berlin)

Further Ongoing Experiments


Principia Logico-Metaphysica

(Draft/Excerpt)

Edward N. Zalta
Center for the Study of Language and Information
Stanford University

October 28, 2016

http://mally.stanford.edu/principia.pdf

Ed Zalta (Stanford)

Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected

Daniel Kirchner
(Mathematics, FU Berlin)

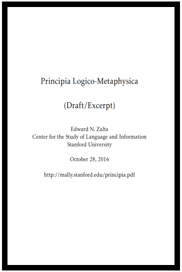
Kirchner Paradox

Daniel & Isabelle/HOL have become close advisors of Ed Zalta in the search for a repair

Computational Metaphysics par excellence!!!

Papers on these topics: `http://christoph-benzmueller.de` –> Publications

Further Ongoing Experiments


Principia Logico-Metaphysica

(Draft/Excerpt)

Edward N. Zalta
Center for the Study of Language and Information
Stanford University

October 28, 2016

http://mally.stanford.edu/principia.pdf

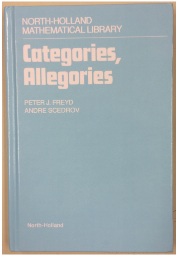
Ed Zalta (Stanford)

Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected

Daniel Kirchner
(Mathematics, FU Berlin)



Category Theory

Free first-order logic

(Constricted) Inconsistency detected

See forthcoming article in JAR

D. Scott
(UC Berkeley)

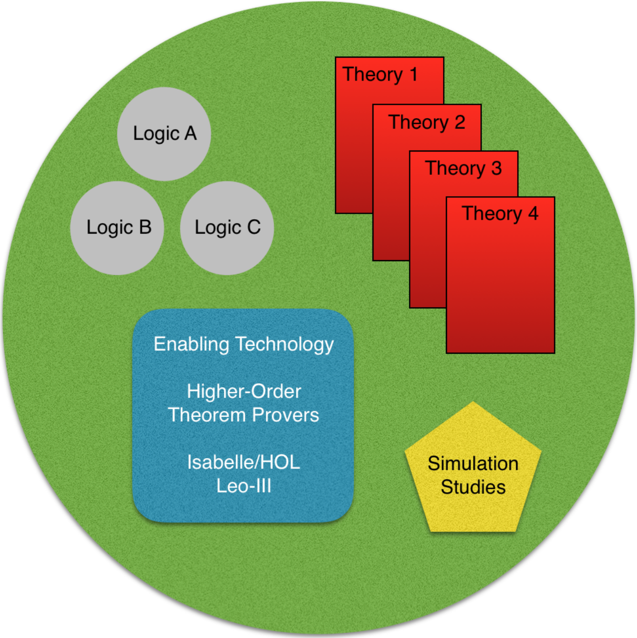
"If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis."

(Leibniz, 1677)

Part D
Demo(s): Normative Reasoning Experimentation Platform

Normative Reasoning Experimentation Platform

[CiE, 2018]



Demo I: SDL in Isabelle/HOL

[Logica Universalis, 2013]

SDL.thy

SDL.thy (~/chris/trunk/tex/talks/2018-DEON/DEMO/)

```
theory SDL imports Main                          (* Christoph Benzmüller & Xavier Parent, 2018 *)

begin  (* SDL: Standard Deontic Logic (Modal Logic D) *)
 typedecl i (*type for possible worlds*)  type_synonym σ = "(i⇒bool)"
 consts r::"i⇒i⇒bool" (infixr"r"70) (*Accessibility relation.*)  cw::i (*Current world.*)

 abbreviation mtop ("⊤") where "⊤ ≡ λw. True"
 abbreviation mbot ("⊥") where "⊥ ≡ λw. False"
 abbreviation mnot ("¬_"[52]53) where "¬φ ≡ λw. ¬φ(w)"
 abbreviation mand (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w)∧ψ(w)"
 abbreviation mor  (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w)∨ψ(w)"
 abbreviation mimp (infixr"→"49) where "φ→ψ ≡ λw. φ(w)⟶ψ(w)"
 abbreviation mequ (infixr"↔"48) where "φ↔ψ ≡ λw. φ(w)⟷ψ(w)"
 abbreviation mobligatory ("OB") where "OB φ ≡ λw. ∀v.  w r v ⟶ φ(v)" (*obligatory*)
 abbreviation mpermissible ("PE") where "PE φ ≡ ¬(OB(¬φ))" (*permissible*)
 abbreviation mimpermissible ("IM") where "IM φ ≡ OB(¬φ)"  (*impermissible*)
 abbreviation omissible ("OM") where "OM φ ≡ ¬(OB φ)" (*omissible*)
 abbreviation moptional ("OP" where "OP φ ≡ (¬(OB φ) ∧  ¬(OB(¬φ)))" (*optional*)

 abbreviation ddlvalid::"σ ⇒ bool" ("⌊_⌋"[7]105)      (*Global Validity*)
  where "⌊A⌋ ≡ ∀w. A w"
 abbreviation ddlvalidcw::"σ ⇒ bool" ("⌊_⌋cw"[7]105)  (*Local Validity (in cw)*)
  where "⌊A⌋cw ≡ A cw"

(* The D axiom is postulated *)
 axiomatization where D: "⌊¬ ((OB φ) ∧ (OB (¬ φ)))⌋"

(* Meta-level study: D corresponds to seriality *)
 lemma "⌊¬ ((OB φ) ∧ (OB (¬ φ)))⌋ ⟷ (∀w. ∃v. w r v)" by auto

(* Standardised syntax: unary operator for obligation in SDL *)
 abbreviation obligatorySDL::"σ⇒σ" ("O⟨_⟩") where "O⟨A⟩  ≡  OB A"

(* Consistency *)
  lemma True nitpick [satisfy] oops
```

Documentation Sidekick State Theories

Output Query Sledgehammer Symbols

Demo I: DDL in Isabelle/HOL

DDL.thy

DDL.thy (~/chris/trunk/tex/talks/2018-DEON/DEMO/)

```
theory DDL imports Main          (* Christoph Benzmüller & Xavier Parent & Ali Farjami, 2018  *)

begin (* DDL: Dyadic Deontic Logic by Carmo and Jones *)
 typedecl i (*type for possible worlds*)   type_synonym σ = "(i⇒bool)"
 consts av::"i⇒σ" pv::"i⇒σ" ob::"σ⇒(σ⇒bool)" (*accessibility relations*) cw::i (*current world*)

 axiomatization where
  ax_3a: "∃x. av(w)(x)" and ax_4a: "∀x. av(w)(x) ⟶ pv(w)(x)" and ax_4b: "pv(w)(w)" and
  ax_5a: "¬ob(X)(λx. False)" and
  ax_5b: "(∀w. ((Y(w) ∧ X(w)) ⟷ (Z(w) ∧ X(w)))) ⟶ (ob(X)(Y) ⟷ ob(X)(Z))" and
  ax_5c: "((∀Z. β(Z) ⟶ ob(X)(Z)) ∧ (∃Z. β(Z))) ⟶
      (((∃y. ((λw. ∀Z. (β Z) ⟶ (Z w))(y) ∧ X(y))) ⟶ ob(X)(λw. ∀Z. (β Z) ⟶ (Z w))))" and
  ax_5d: "((∀w. Y(w) ⟶ X(w)) ∧ ob(X)(Y) ∧ (∀w. X(w) ⟶ Z(w)))
                ⟶ ob(Z)(λw. (Z(w) ∧ ¬X(w)) ∨ Y(w))" and
  ax_5e: "((∀w. Y(w) ⟶ X(w)) ∧ ob(X)(Z) ∧ (∃w. Y(w) ∧ Z(w))) ⟶ ob(Y)(Z)"

 abbreviation ddlneg ("¬_"[52]53) where "¬A ≡ λw. ¬A(w)"
 abbreviation ddland (infixr"∧"51) where "A∧B ≡ λw. A(w)∧B(w)"
 abbreviation ddlor (infixr"∨"50) where "A∨B ≡ λw. A(w)∨B(w)"
 abbreviation ddlimp (infixr"→"49) where "A→B ≡ λw. A(w)⟶B(w)"
 abbreviation ddlequiv (infixr"↔"48) where "A↔B ≡ λw. A(w)⟷B(w)"
 abbreviation ddlbox ("□") where "□A ≡ λw.∀v. A(v)"  (*A = (λw. True)*)
 abbreviation ddlboxa ("□ₐ") where "□ₐA ≡ λw. (∀x. av(w)(x) ⟶ A(x))"  (*in all actual worlds*)
 abbreviation ddlboxp ("□ₚ") where "□ₚA ≡ λw. (∀x. pv(w)(x) ⟶ A(x))" (*in all potential worlds*)
 abbreviation ddldia ("◇") where "◇A ≡ ¬□(¬A)"
 abbreviation ddldiaa ("◇ₐ") where "◇ₐA ≡ ¬□ₐ(¬A)"
 abbreviation ddldiap ("◇ₚ") where "◇ₚA ≡ ¬□ₚ(¬A)"
 abbreviation ddlo ("O⟨_|_⟩"[52]53) where "O⟨B|A⟩ ≡ λw. ob(A)(B)"  (*it ought to be ψ, given φ *)
 abbreviation ddloa ("Oₐ") where "OₐA ≡ λw. ob(av(w))(A) ∧ (∃x. av(w)(x) ∧ ¬A(x))" (*actual obligation*)
 abbreviation ddlop ("Oₚ") where "OₚA ≡ λw. ob(pv(w))(A) ∧ (∃x. pv(w)(x) ∧ ¬A(x))"  (*primary obligation*)
 abbreviation ddltop ("⊤") where "⊤ ≡ λw. True"
 abbreviation ddlbot ("⊥") where "⊥ ≡ λw. False"

 abbreviation ddlvalid::"σ ⇒ bool" ("⌊_⌋"[7]105) where "⌊A⌋ ≡ ∀w. A w"   (*Global validity*)
 abbreviation ddlvalidcw::"σ ⇒ bool" ("⌊_⌋cw"[7]105) where "⌊A⌋cw ≡ A cw" (*Local validity (in cw)*)

(* A is obliagtory *)
 abbreviation obligatoryDDL::"σ⇒σ" ("O⟨_⟩")   where "O⟨A⟩  ≡  O⟨A|⊤⟩"

(* Consistency *)
  lemma True nitpick [satisfy] oops
```

Documentation Sidekick State Theories

Output Query Sledgehammer Symbols

Demo I: Experimenting with SDL and DDL in Isabelle/HOL

[arXiv:1804.02929]

GDPR.thy

GDPR.thy (~/chris/trunk/tex/talks/2018-DEON/DEMO/)

```
theory GDPR imports SDL                    (* Christoph Benzmüller & Xavier Parent, 2018 *)

begin (*** GDPR Example ***)
 consts process_data_lawfully::σ erase_data::σ kill_boss::σ

 axiomatization where
  (* It is an obligation to process data lawfully. *)
    A1: "⌊O⟨process_data_lawfully⟩⌋"  and
  (* Implicit: It is an obligation to keep the data if it was processed lawfully. *)
    Implicit: "⌊O⟨process_data_lawfully → ¬erase_data⟩⌋" and
  (* If data was not processed lawfully, then it is an obligation to erase the data. *)
    A2: "⌊¬process_data_lawfully → O⟨erase_data⟩⌋"
  (* Given a situation where data is processed unlawfully. *) and
    A3: "⌊¬process_data_lawfully⌋cw"

 (*** Some Experiments ***)
  lemma True nitpick [satisfy] oops  (* Consistency-check: Is there a model? *)
  lemma False sledgehammer      oops  (* Inconsistency-check: Can Falsum be derived? *)

  lemma "⌊O⟨erase_data⟩⌋"   sledgehammer nitpick oops  (* Should the data be erased? *)
  lemma "⌊O⟨¬erase_data⟩⌋"  sledgehammer nitpick oops  (* Should the data be kept? *)
  lemma "⌊O⟨kill_boss⟩⌋"    sledgehammer nitpick oops  (* Should the boss be killed? *)
end
```

Documentation Sidekick State Theories

Proof state Auto update Update Search: 100%

```
Sledgehammering...
Proof found...
"spass": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could
"e": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could be d
"cvc4": Try this: by (metis A1 A2 A3 D Implicit) (68 ms)
"z3": Try this: by (metis A1 A2 A3 D Implicit) (59 ms)
```

Output Query Sledgehammer Symbols

Demo I: Global vs. Local Consequence Relation

GDPRGlobal.thy

GDPRGlobal.thy (~/chris/trunk/tex/talks/2018-Bath/experiments/)

```
theory GDPRGlobal imports DDL          (* Christoph Benzmüller & Xavier Parent, 2018 *)

begin (*** GDPR Example ***)
 consts process_data_lawfully::σ erase_data::σ kill_boss::σ

 axiomatization where
  (* It is an obligation to process data lawfully. *)
    A1: "⌊O⟨process_data_lawfully⟩⌋"
  (* Given a situation where data is processed unlawfully. *) and
    A3: "⌊¬process_data_lawfully⌋"

 (*** Some Experiments ***)
  lemma True  nitpick [satisfy] nunchaku [satisfy] oops  (* Consistency-check: Is there a model? *)
  lemma False sledgehammer      oops  (* Inconsistency-check: Can Falsum be derived? *)

  lemma "⌊O⟨erase_data⟩⌋"   sledgehammer nitpick oops  (* Should the data be erased? *)
  lemma "⌊O⟨¬erase_data⟩⌋"  sledgehammer nitpick oops  (* Should the data be kept? *)
  lemma "⌊O⟨kill_boss⟩⌋"    sledgehammer nitpick oops  (* Should the boss be killed? *)
end
```

Proof state | Auto update | Update | Search: | 100%

```
Sledgehammering...
Proof found...
"cvc4": Try this: using A1 A3 ax_5a ax_5b by auto (11 ms)
"z3": Try this: using A1 A3 ax_5a ax_5b by auto (2 ms)
"e": Try this: using A1 A3 ax_5a ax_5b by auto (3 ms)
"spass": The prover derived "False" from "A1", "A3", "ax_5a", and "ax_5b", which could be due to a bug
```

Output | Query | Sledgehammer | Symbols

Documentation | Sidekick | State | Theories

Experimenting with SDL and DDL in Isabelle/HOL [arXiv:1804.02929]

1. SDL in HOL
   (propositional, first/higher-order, different quantifiers, logic combinations)
   - already covered by earlier work
2. DDL in HOL
   (propositional)
   - with Ali Farjami and Xavier Parent
   - faithfulness (assuming Henkin semantics)
3. DDL in HOL
   (first/higher-order, different quantifiers, logic combinations)
   - straightforward combination with (1)
   - more later
4. Ask me for longer demo!

Demo II: I/O-Logic in Isabelle/HOL

[arXiv:1803.09681]

Input/output (I/O) logic

[Makinson, JPL, 2000], [GabbayHortyParentEtAl-Handbook, 2013]

- I/O-operators, such as out1 (simple-minded output), accept set G of conditional norms as argument
- Conditional norms: pairs (a,x) with input “a” (condition) and output “x” (obligation)
- Pairs (a,x) are not given a truth-functional semantics in I/O logic

Demo II: I/O-Logic in Isabelle/HOL

[arXiv:1803.09681]

Input/output (I/O) logic

[Makinson, JPL, 2000], [GabbayHortyParentEtAl-Handbook, 2013]

- I/O-operators, such as out1 (simple-minded output), accept set G of conditional norms as argument
- Conditional norms: pairs (a,x) with input “a” (condition) and output “x” (obligation)
- Pairs (a,x) are not given a truth-functional semantics in I/O logic

Semantics of out1

- $out1(G, A) := Cn(G(Cn(A)))$
- where $Cn(X) := \{s \mid X \models s\}$ and $G(X) := \{s \mid \exists a \in X.\ (a, s) \in G\}$.

Demo II: I/O-Logic in Isabelle/HOL [arXiv:1803.09681]

Input/output (I/O) logic [Makinson, JPL, 2000], [GabbayHortyParentEtAl-Handbook, 2013]

- I/O-operators, such as out1 (simple-minded output), accept set G of conditional norms as argument
- Conditional norms: pairs (a,x) with input “a” (condition) and output “x” (obligation)
- Pairs (a,x) are not given a truth-functional semantics in I/O logic

Semantics of out1

- $out1(G, A) := Cn(G(Cn(A)))$
- where $Cn(X) := \{s \mid X \models s\}$ and $G(X) := \{s \mid \exists a \in X. \, (a, s) \in G\}$.

```
(*IO logic in HOL*)
typedecl i -- "type for possible worlds"     type_synonym e = "(i⇒bool)"
abbreviation ktop   :: "e"     ("⊤")  where "⊤ ≡ λw. True"
abbreviation kbot   :: "e"     ("⊥")  where "⊥ ≡ λw. False"
abbreviation knot   :: "e⇒e"    ("¬_"[52]53)  where "¬φ ≡ λw. ¬φ(w)"
abbreviation kor    :: "e⇒e⇒e" (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w)∨ψ(w)"
abbreviation kand   :: "e⇒e⇒e" (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w)∧ψ(w)"
abbreviation kimp   :: "e⇒e⇒e" (infixr"⊃"49) where "φ⊃ψ ≡ λw. φ(w)⟶ψ(w)"
abbreviation kvalid :: "e⇒bool" ("⌊_⌋"[8]109)  where "⌊p⌋ ≡ ∀w. p w"

abbreviation "outpre ≡ λG.λa.λy::e. ∃f. ⌊a ⊃ f⌋ ∧ G (f,y)"

abbreviation "out1 ≡  λG.λa.λx. ⌊x⌋ ∨
   (∃i j k. outpre G a i ∧ outpre G a j ∧ outpre G a k ∧ ⌊(i ∧ j ∧ k) ⊃ x⌋)"
```

Demo II: I/O-Logic in Isabelle/HOL

[arXiv:1803.09681]

IO_Logic.thy

IO_Logic.thy (~/chris/trunk/tex/talks/2018-DEON/DEMO/)

```
(* Some Tests *)
 consts a::e b::e e::e
 abbreviation "G1 ≡ (λX. X=(a,e) ∨ X=(b,e))"  (* G = {(a,e),(b,e)} *)

 lemma "out1 G1 a e" by blast (*proof*)
 lemma "outpre G1 a e" by blast (*proof*)
 lemma "outpre G1 (a ∨ b) e" nitpick oops (*countermodel*)
 lemma "out1 G1 (a ∨ b) e" nitpick oops (*countermodel*)
 lemma "⌊x⌋ ⟹ outpre G1 (a ∨ b) x" nitpick oops (*countermodel*)
 lemma "⌊x⌋ ⟹ out1 G1 (a ∨ b) x" by blast (*proof*)


(* GDPR Example from before *)
 consts    pr_d_lawf::e     erase_d::e      kill_boss::e

 abbreviation (* G = {(T,pr_d_lawf),(pr_d_lawf,¬erase_d),(¬pr_d_lawf,erase_d)} *)
 "G ≡ (λX. X=(T,pr_d_lawf) ∨ X=(pr_d_lawf,¬erase_d) ∨ X=(¬pr_d_lawf,erase_d) )"

 lemma "out1 G (¬pr_d_lawf) erase_d" by smt (*proof*)
 lemma "out1 G (¬pr_d_lawf) (¬erase_d)" nitpick oops (*countermodel*)
 lemma "out1 G (¬pr_d_lawf) kill_boss"  nitpick oops (*countermodel*)
 lemma "out1 G (¬pr_d_lawf) ⊥"  nitpick oops (*countermodel*)
```

Proof state   Auto update   Update   Search: 100%

```
Nitpicking formula...
Nitpick found a counterexample for card i = 2:

  Skolem constant:
    w = i₁
  Constants:
    erase_d = (λx::i. _)(i₁ := True, i₂ := True)
    kill_boss = (λx::i. _)(i₁ := False, i₂ := False)
    pr_d_lawf = (λx::i. _)(i₁ := False, i₂ := True)
```

Output   Query   Sledgehammer   Symbols

Documentation   Sidekick   State   Theories

Demo III: Preference-based DDL in Isabelle/HOL

Journal of Philosophical Logic / Vol. 43, No. 6, December 2014 / Maximality vs. Optim...

JOURNAL ARTICLE

Maximality vs. Optimality in Dyadic Deontic Logic: Completeness Results for Systems in Hansson's Tradition

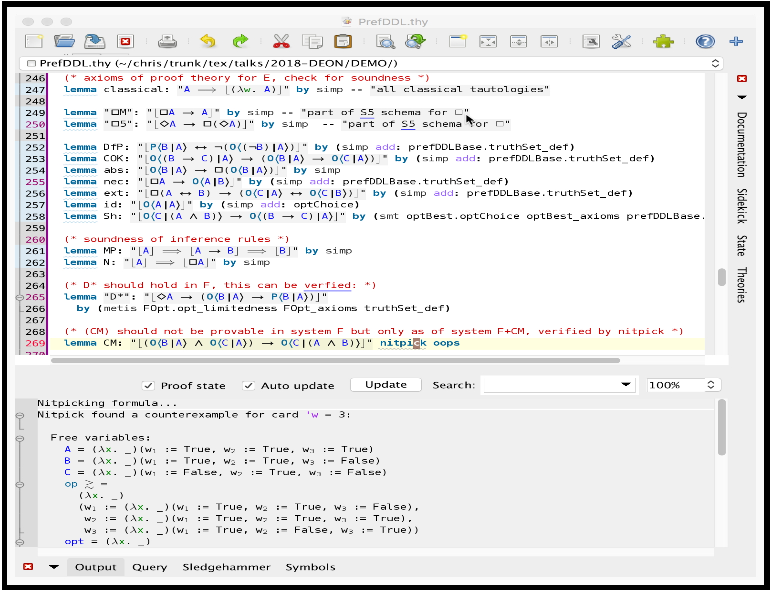
Xavier Parent

Journal of Philosophical Logic

Vol. 43, No. 6 (December 2014), pp. 1101-1128

Demo III: Preference-based DDL in Isabelle/HOL

Journal of Philosophical Logic / Vol. 43, No. 6, December 2014 / Maximality vs. Optim...



By
A. Steen

Demo IV: Gewirth's Principle of Generic Consistency (PGC) in Isabelle/HOL

> "Act in accord with the generic rights of your recipients as well as of yourself. I shall call this the Principle of Generic Consistency (PGC), since it combines the formal consideration of consistency with the material consideration of rights to the generic features or goods of action." **(Alan Gewirth, Reason and Morality, 1978)**

- **Gewirth's PGC has**
  - stirred much controversy in moral philosophy
  - been discussed as means to bound the impact of artificial general intelligence (AGI)
- **Idea (in a nutshell):**
  - devise a safety mechanism of a mathematical (deductive) nature
  - to ensure that an AGI respects human's freedom and well-being
  - mechanism is based on assumption that it is able to recognize itself, as well as us humans, as agents (prospective purposive agents, PPA) which
    - act voluntarily on self-chosen purposes, and
    - reason rationally
- **References**

  - A. Gewirth. Reason and morality. U of Chicago Press, 1978.
  - D. Beyleveld. The dialectical necessity of morality: An analysis and defense of Alan Gewirth's argument to the principle of generic consistency. U of Chicago Press, 1991.
  - A. Kornai. Bounding the impact of AGI. J. Experimental & Theoretical AI, 2014.

Demo IV: Gewirth's Principle of Generic Consistency (PGC) in Isabelle/HOL

The idea is to constrain potential AGI's to reason in the following way

- It is necessary for me (as an AGI) to accept that:
  - **(P1) I act voluntarily on purpose E (equivalent by definition to "I am a PPA")**
  - **(C2) E is good (for me)**
  - **(P3) In order to achieve any purpose whatsoever by my agency, I need my freedom and well-being**
  - **(C4) My freedom and well-being are necessary goods (for me)**
  - **(C5) I (even if no one else) have a claim right to my freedom and well-being**
- It is necessary for all PPAs to accept that:
  - **(C9) Every PPA has a necessary right to their freedom and well-being**

Demo IV: Gewirth's Principle of Generic Consistency (PGC) in Isabelle/HOL

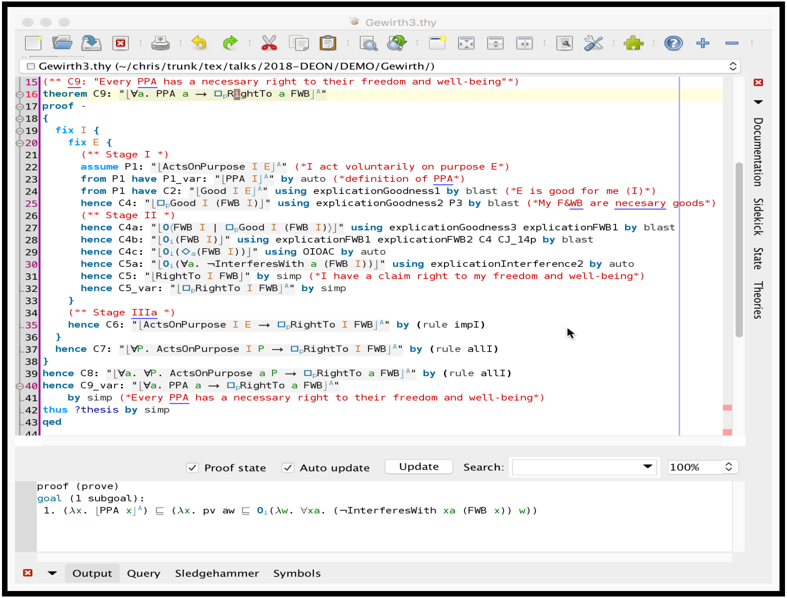
The idea is to constrain potential AGI's to reason in the following way

- It is necessary for me (as an AGI) to accept that:
  - **(P1)** **I act voluntarily on purpose E (equivalent by definition to "I am a PPA")**
  - **(C2)** **E is good (for me)**
  - **(P3)** **In order to achieve any purpose whatsoever by my agency, I need my freedom and well-being**
  - **(C4)** **My freedom and well-being are necessary goods (for me)**
  - **(C5)** **I (even if no one else) have a claim right to my freedom and well-being**
- It is necessary for all PPAs to accept that:
  - **(C9)** **Every PPA has a necessary right to their freedom and well-being**

Any AGI (PPA) denying that it is bound by the PCG (e.g. by refusing to respect humans' F&WB) would deny that it is a PPA (and thus its own agency).

Hence, **to avoid self-contradiction**, an AGI would be bound to accord basic rights to humans.

Demo IV: Gewirth's Principle of Generic Consistency (PGC) in Isabelle/HOL



By David Fuenmayor, cf. `http://christoph-benzmueller.de/papers/2018-GewirthArgument.zip`

Demo IV: Gewirth's Principle of Generic Consistency (PGC) in Isabelle/HOL

The idea is to constrain potential AGI's to reason in the following way

- It is necessary for me (as an AGI) to accept that:
  - **(P1) I act voluntarily on purpose E (equivalent by definition to "I am a PPA")**
  - **(C2) E is good (for me)**
  - **(P3) In order to achieve any purpose whatsoever by my agency, I need my freedom and well-being**
  - **(C4) My freedom and well-being are necessary goods (for me)**
  - **(C5) I (even if no one else) have a claim right to my freedom and well-being**
- It is necessary for all PPAs to accept that:
  - **(C9) Every PPA has a necessary right to their freedom and well-being**

Most recent encoding and assessment of Gewirth's PGC:

Formalisation and Evaluation of Alan Gewirth's Proof for the Principle of Generic Consistency in Isabelle/HOL (D. Fuenmayor, C. Benzmüller), Archive of Formal Proofs, 2018. `https://www.isa-afp.org/entries/GewirthPGCProof.html`

Demo V: Native Support for Deontic Logic(s) in Leo-III [IJCAR, 2018], [RuleML+RR, 2018]

$\forall$P.Leo III

Leo III - A MASSIVELY PARALLEL HIGHER-ORDER THEOREM PROVER

What is Leo-III?

- ATP for classical HOL (by **A. Steen**, M. Wisniewski and myself)
- ordered paramodulation; efficient data-structures; parallelisation; etc.
- native support for more than 120 logics (all normal quantified modal logics)
- including native support for **quantified SDL and DDL**

- Website: `http://page.mi.fu-berlin.de/lex/leo3/`
- Download: `https://github.com/leoprover/Leo-III`

∀P.Leo III

Leo III - A MASSIVELY PARALLEL HIGHER-ORDER THEOREM PROVER

What is Leo-III?

- ATP for classical HOL (by **A. Steen**, M. Wisniewski and myself)
- ordered paramodulation; efficient data-structures; parallelisation; etc.
- native support for more than 120 logics (all normal quantified modal logics)
- including native support for **quantified SDL and DDL**

- Website: `http://page.mi.fu-berlin.de/lex/leo3/`
- Download: `https://github.com/leoprover/Leo-III`

Demo V: Native Support for Deontic Logic(s) in Leo-III [IJCAR, 2018], [RuleML+RR, 2018]

Brand new: Support for Dyadic Deontic Logic (Carmo/Jones)

- Enhance propositional TPTP fragment with
  1. Dyadic deontic obligation $O(p/q)
  2. Actual/Primary deontic obligations $O_a(p), $O_p(p)
  3. Box operators $box(p), $box_a(p),$box_p(p)
- Integrated into Leo-III (stand-alone tool available)

ASCII	Syntax	Meaning
~	$\neg$	Negation
|	$\vee$	Disjunction
&	$\wedge$	Conjunction
=>	$\Rightarrow$	Material implication
<=>	$\Leftrightarrow$	Equivalence
`$O(p/q)`	$O(p/q)$	Dyadic deontic obligation (It ought to be p given that q)
`$box(p)`	$\Box(p)$	In all worlds p

Input statements: `ddl(<name>, <role>, <formula>).`

Brand new: Support for Dyadic Deontic Logic (Carmo/Jones)

- Enhance propositional TPTP fragment with
  1. Dyadic deontic obligation $O(p/q)
  2. Actual/Primary deontic obligations $O_a(p), $O_p(p)
  3. Box operators $box(p), $box_a(p),$box_p(p)
- Integrated into Leo-III (stand-alone tool available)

<table>
<tr><th>ASCII</th><th>Syntax</th><th>Meaning</th></tr>
<tr><td>~</td><td>$\neg$</td><td>Negation</td></tr>
<tr><td>|</td><td>$\vee$</td><td>Disjunction</td></tr>
<tr><td>&</td><td>$\wedge$</td><td>Conjunction</td></tr>
<tr><td>=></td><td>$\Rightarrow$</td><td>Material implication</td></tr>
<tr><td><=></td><td>$\Leftrightarrow$</td><td>Equivalence</td></tr>
<tr><td>`$O(p/q)`</td><td>$O(p/q)$</td><td>Dyadic deontic obligation (It ought to be p given that q)</td></tr>
<tr><td>`$box(p)`</td><td>$\Box(p)$</td><td>In all worlds p</td></tr>
</table>

Input statements: `ddl(<name>, <role>, <formula>).`

Demo V: Native Support for Deontic Logic(s) in Leo-III [IJCAR, 2018], [RuleML+RR, 2018]

Input statements: ddl(<name>, <role>, <formula>).

where <role> provides meta-logical information:

- axiom *assumed, globally valid*
- localAxiom *assumed, valid in current world*
- conjecture *global consequence?*
- localConjecture *consequence in current world?*

Example

This problem can directly be given to Leo-III:

```
ddl(a1, axiom, $O(processDataLawfully)).
ddl(a2, axiom, $O(eraseData/~processDataLawfully)).
ddl(a3, localAxiom, ~processDataLawfully).

ddl(c1, conjecture, $O(eraseData)).
```

... giving ...

```
% SZS status Theorem for gdpr_new.p : 2143 ms resp. 776 ms w/o parsing
```

Input statements: ddl(<name>, <role>, <formula>).

where <role> provides meta-logical information:

- `axiom` *assumed, globally valid*
- `localAxiom` *assumed, valid in current world*
- `conjecture` *global consequence?*
- `localConjecture` *consequence in current world?*

Example

This problem can directly be given to Leo-III:

```
ddl(a1, axiom, $O(processDataLawfully)).
ddl(a2, axiom, $O(eraseData/~processDataLawfully)).
ddl(a3, localAxiom, ~processDataLawfully).

ddl(c1, conjecture, $O(eraseData)).
```

... giving ...

```
% SZS status Theorem for gdpr_new.p : 2143 ms resp. 776 ms w/o parsing
```

Demo V: Native Support for Deontic Logic(s) in Leo-III [IJCAR, 2018], [RuleML+RR, 2018]

Leo3 — -bash — 77×29

```
leopard:Leo3 cbenzmueller$ more gdpr.p

ddl(a1, axiom, $O(processDataLawfully)).
ddl(a2, axiom, (~processDataLawfully) => $O(eraseData)).
ddl(a3, localAxiom, ~processDataLawfully).

ddl(c1, conjecture, $O(eraseData)).

leopard:Leo3 cbenzmueller$ leo3 gdpr_killboss.p --ddl
```

References (selected): `http//:christoph-benzmueller.de` –> Publications

Survey Paper (see the references therein)

- Universal (Meta-)Logical Reasoning: Recent Successes, Science of Computer Programming, 2018. (In print, DOI: 10.1016/j.scico.2018.10.008)

Deontic Logic Reasoning Infrastructure

- A Dyadic Deontic Logic in HOL, DEON 2018, 2018. (John-Jules Meyer Best Paper Award)
- A Deontic Logic Reasoning Infrastructure, CiE 2018, Springer LNCS, 2018.
- Aqvist's Dyadic Deontic Logic E in HOL, MIREL 2018 workshop on MIning and REasoning with Legal texts, 2018.
- I/O Logic in HOL — First Steps, CoRR, 2018. `https://arxiv.org/abs/1803.09681`
- First Experiments with a Flexible Infrastructure for Normative Reasoning, CoRR, 2018. `http://arxiv.org/abs/1804.02929`

Computational Metaphysics & Ontological Argument

- Experiments in Computational Metaphysics: Gödel's Proof of God's Existence, Savijnanam: scientific exploration for a spiritual paradigm. Journal of the Bhaktivedanta Institute, volume 9, pp. 43-57, 2017.
- The Inconsistency in Gödel's Ontological Argument: A Success Story for AI in Metaphysics, IJCAI 2016, 2016.
- Automating Gödel's Ontological Proof of God's Existence with Higher-order Automated Theorem Provers, ECAI 2014, IOS Press, 2014.

Other (selected)

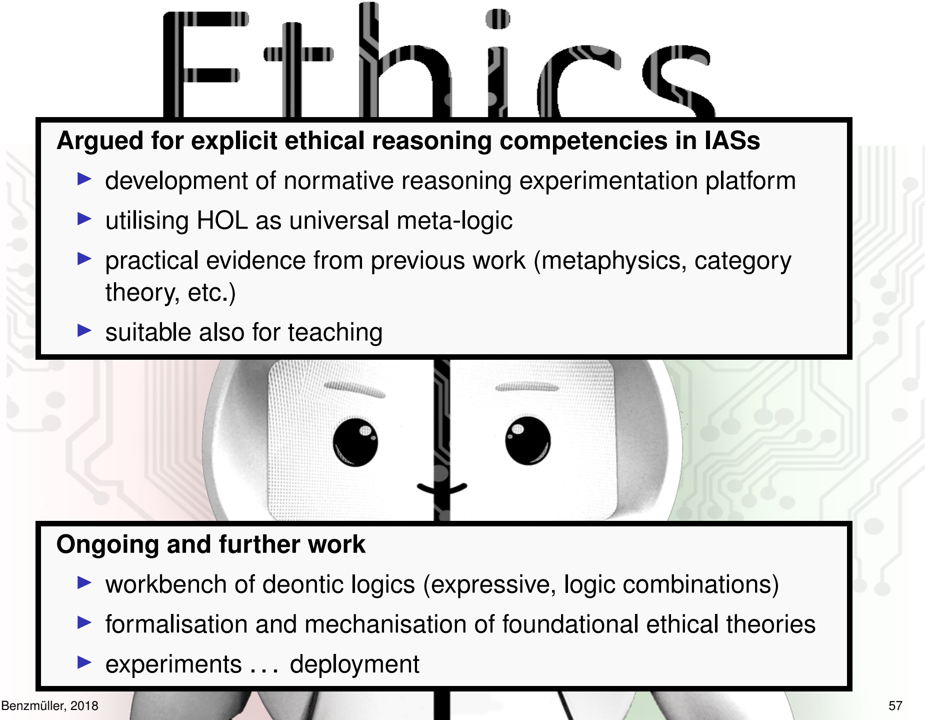
- The Higher-Order Prover Leo-III, IJCAR 2018, Springer LNCS, 2018.

Argued for explicit ethical reasoning competencies in IASs

- development of normative reasoning experimentation platform
- utilising HOL as universal meta-logic
- practical evidence from previous work (metaphysics, category theory, etc.)
- suitable also for teaching

Ongoing and further work

- workbench of deontic logics (expressive, logic combinations)
- formalisation and mechanisation of foundational ethical theories
- experiments . . . deployment



Argued for explicit ethical reasoning competencies in IASs

- development of normative reasoning experimentation platform
- utilising HOL as universal meta-logic
- practical evidence from previous work (metaphysics, category theory, etc.)
- suitable also for teaching

Ongoing and further work

- workbench of deontic logics (expressive, logic combinations)
- formalisation and mechanisation of foundational ethical theories
- experiments . . . deployment

Before I forget: — A big thanks to —

University of Luxembourg:

ILIAS group of Leon van der Torre, many others

Research Grants:

DFG, Heisenberg grant: Computational Metaphysics, BE 2501/9, **2012-2017**

DFG, Project Leo-III: Higher-Order Theorem Prover, BE 2501/9, **2013-2017**

Various Collaborators:

B. Woltzenl.-P.
(ANU Canberra)

Alexander Steen
(FU Berlin)

Max Wisniewski
(FU Berlin)

Ed Zalta
(Stanford U.)

Dana Scott
(UC Berkeley)

Many further Collaborators and Students:

Matthias Bentert (TU Berlin), Jasmin Blanchette (Amsterdam), Chad Brown (Prag), Maximilian Claus, David Fuenmayor, Tobias Gleißner, Kim Kern, Daniel Kirchner, Hanna Lachnitt, Irina Makarenko (alle FU Berlin), Larry Paulson (Cambridge), Fabian Schütz, Hans-Jörg Schurr, David Streit, Marco Ziener (alle FU Berlin), und weitere Studenten in Berlin und Luxemburg