

Studies in Computational Metaphysics & Computational (Pseudo-)Ethics

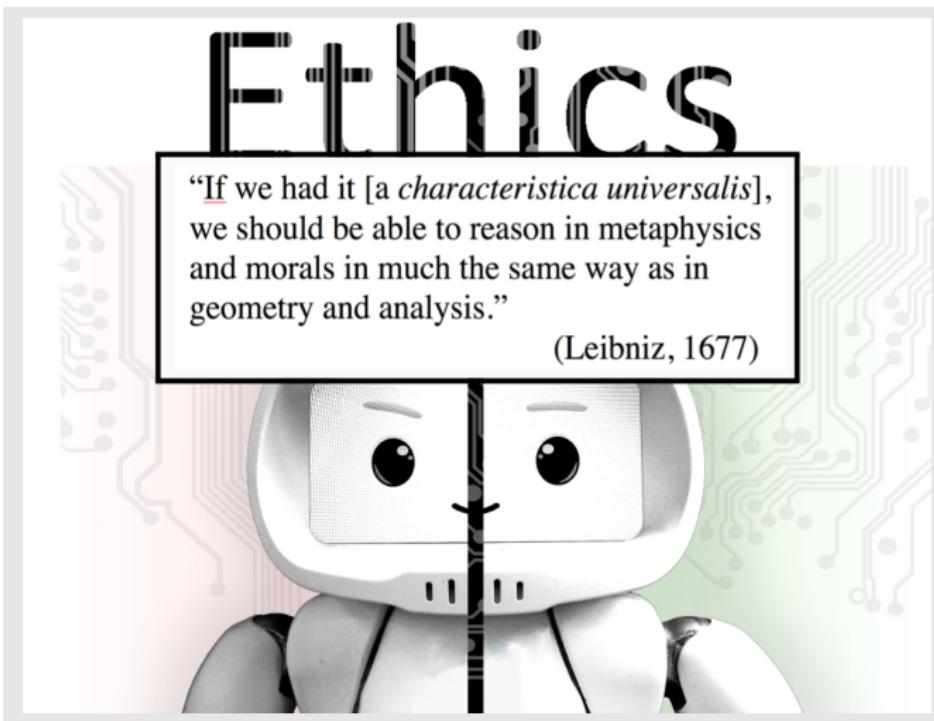
Christoph Benzmüller

Freie Universität Berlin | University of Luxembourg

Ethics

*"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)

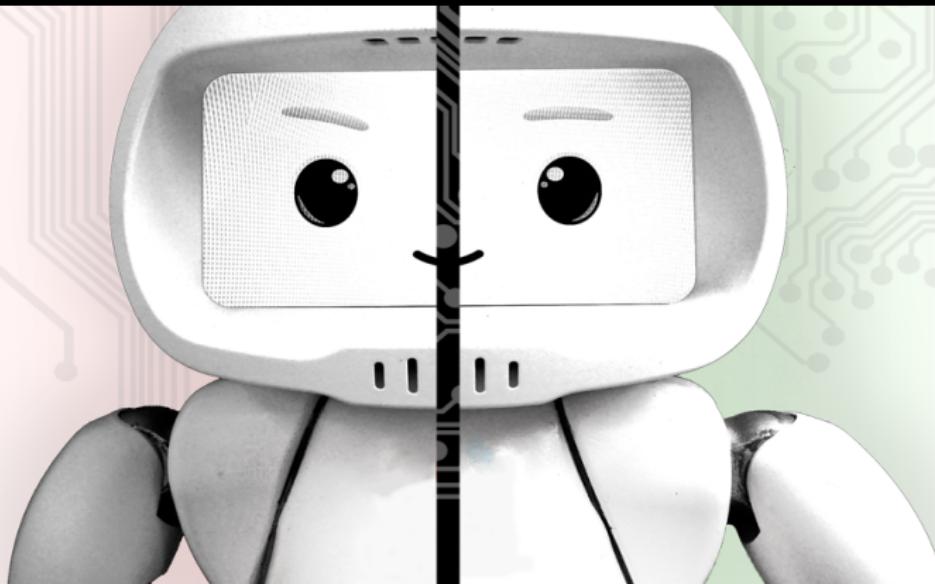


Colloquium Cognitive Systems, U Ulm, 29 Nov 2018

Ethics

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

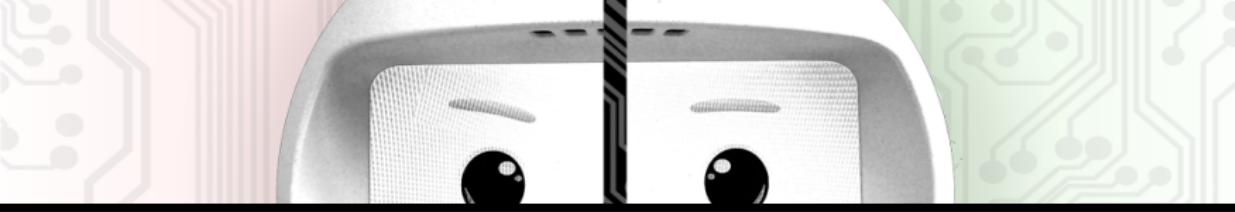
- ▶ appropriate forms of **machine-control**
- ▶ appropriate forms of **human-machine-interaction**



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:

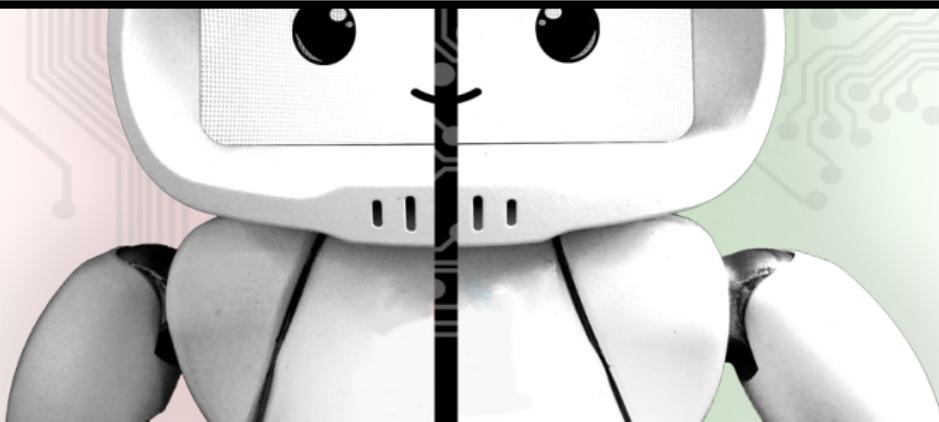
- ▶ **rational argumentation & dialog**
- ▶ **explicit normative reasoning** (legal & ethical)

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

Long-term: Emerging Superintelligence

Really? Anyhow ...

- ▶ How to prevent Superintelligence from turning against humanity?

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

- ▶ 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 IAs

- ▶ 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 IAs

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

Different kinds of systems and approaches:

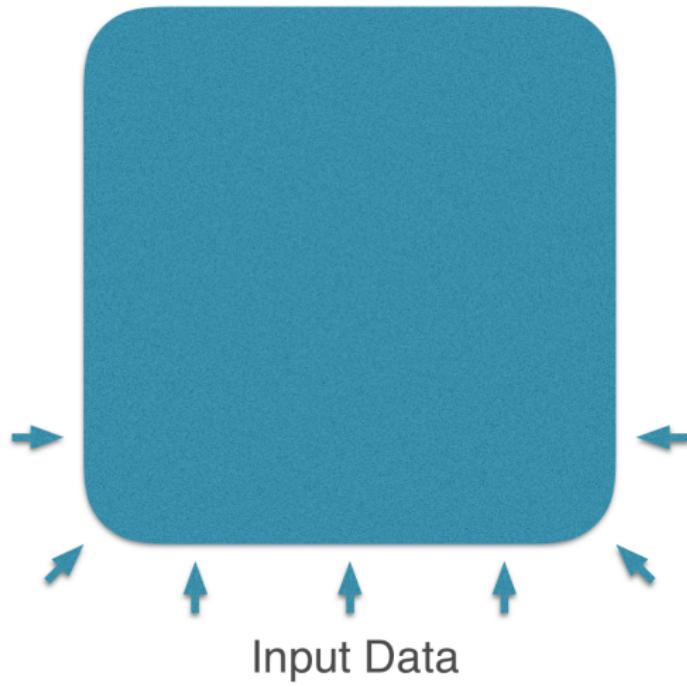
- ▶ [Moor, 2009]:
 - ethical impact agents (ethical consequences to actions)
 - implicit ethical agents (ethical reactions to given situations)
 - **explicit ethical agents** (reasoning with ethical theories/rules)
 - full ethical agents (conscious, intentional, free will)
- ▶ bottom-up vs. **top-down**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

Pseudo-Ethical IAS (medium-term)

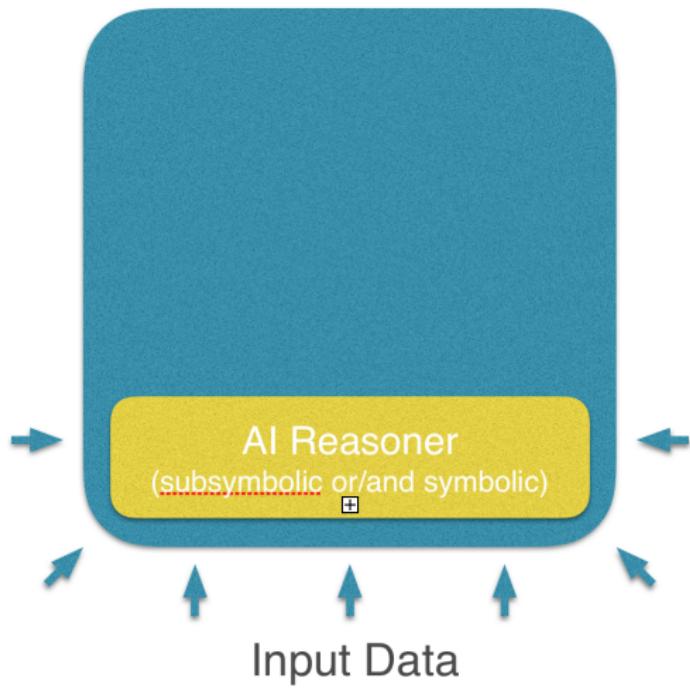


IAS

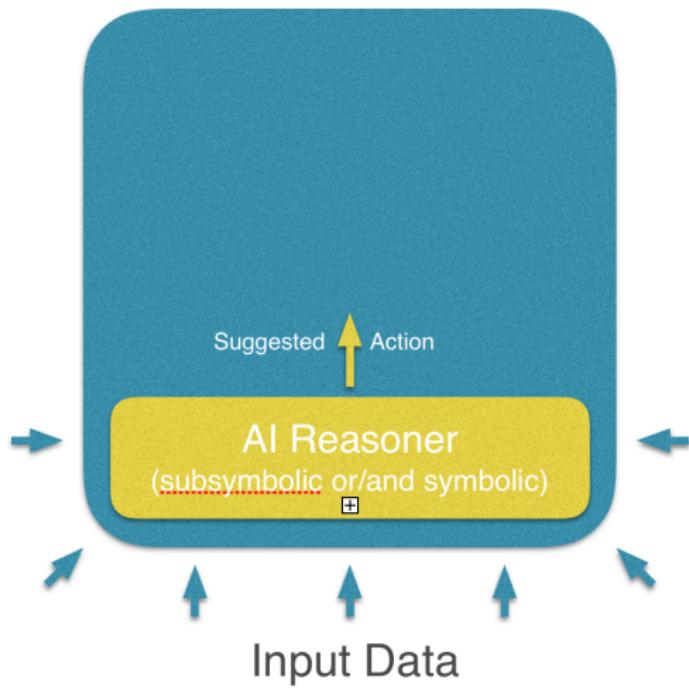
Pseudo-Ethical IAS (medium-term)



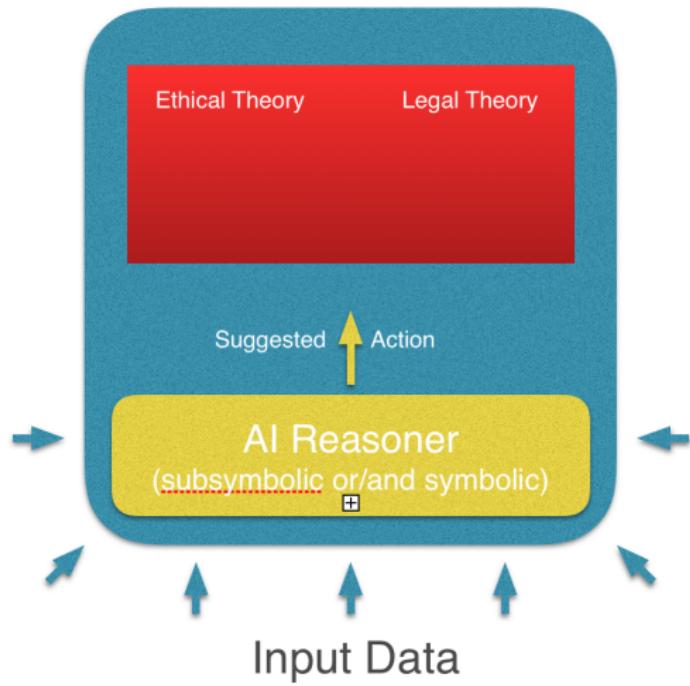
Pseudo-Ethical IAS (medium-term)



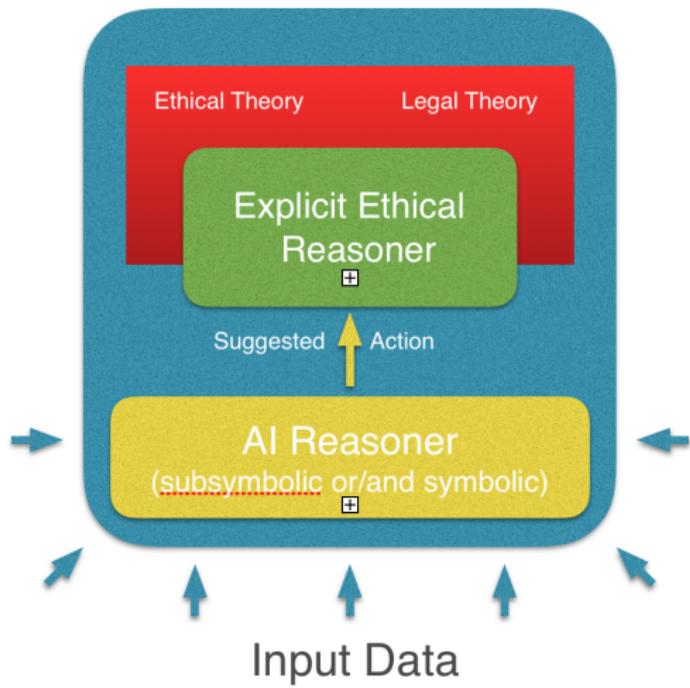
Pseudo-Ethical IAS (medium-term)



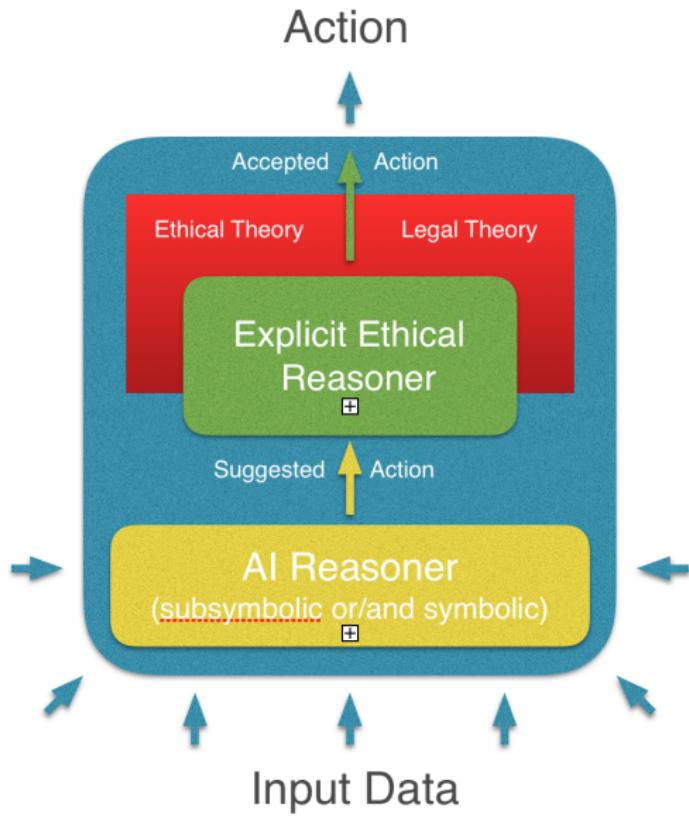
Pseudo-Ethical IAS (medium-term)



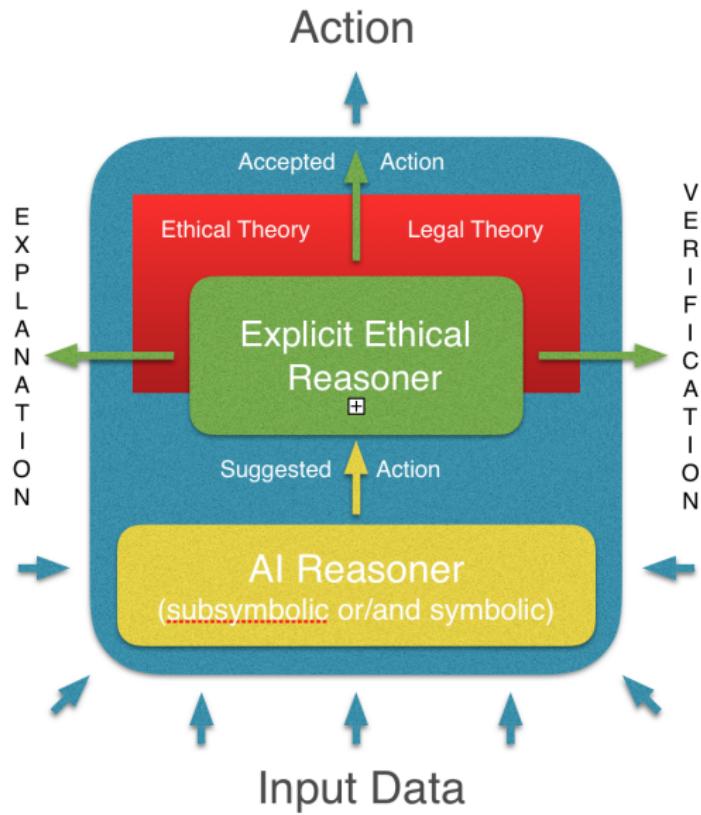
Pseudo-Ethical IAS (medium-term)



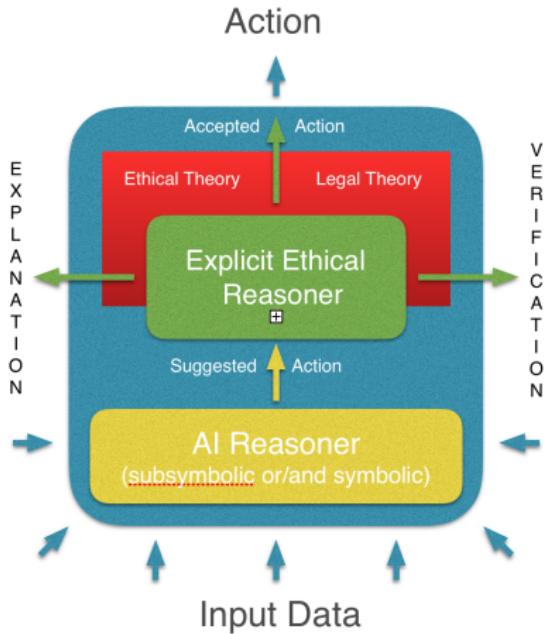
Pseudo-Ethical IAS (medium-term)



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

Deontic Logic

- ▶ Reasoning about obligations and permissions
- ▶ Two groups of approaches:
 - Possible worlds
 - ▶ standard deontic logic
 - ▶ dyadic deontic logic
 - Norm-based semantics
 - ▶ input/output logic

CTD: no

CTD: yes

CTD: yes

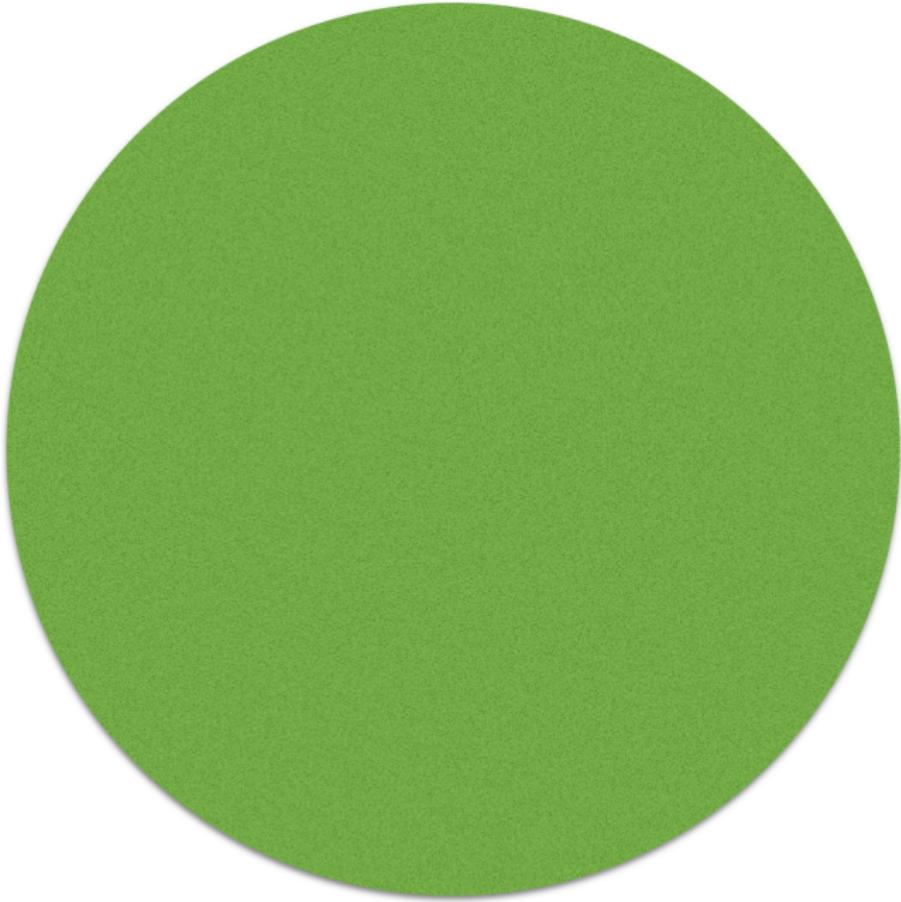


A. Farjam

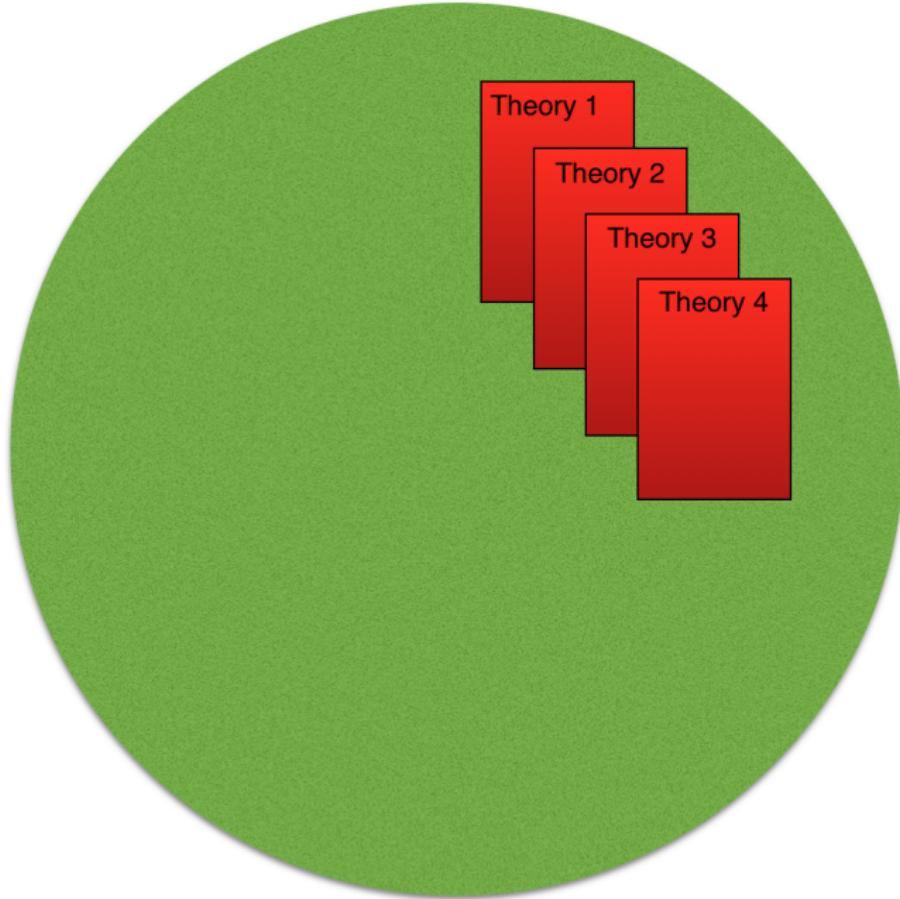
Further interests and challenges

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

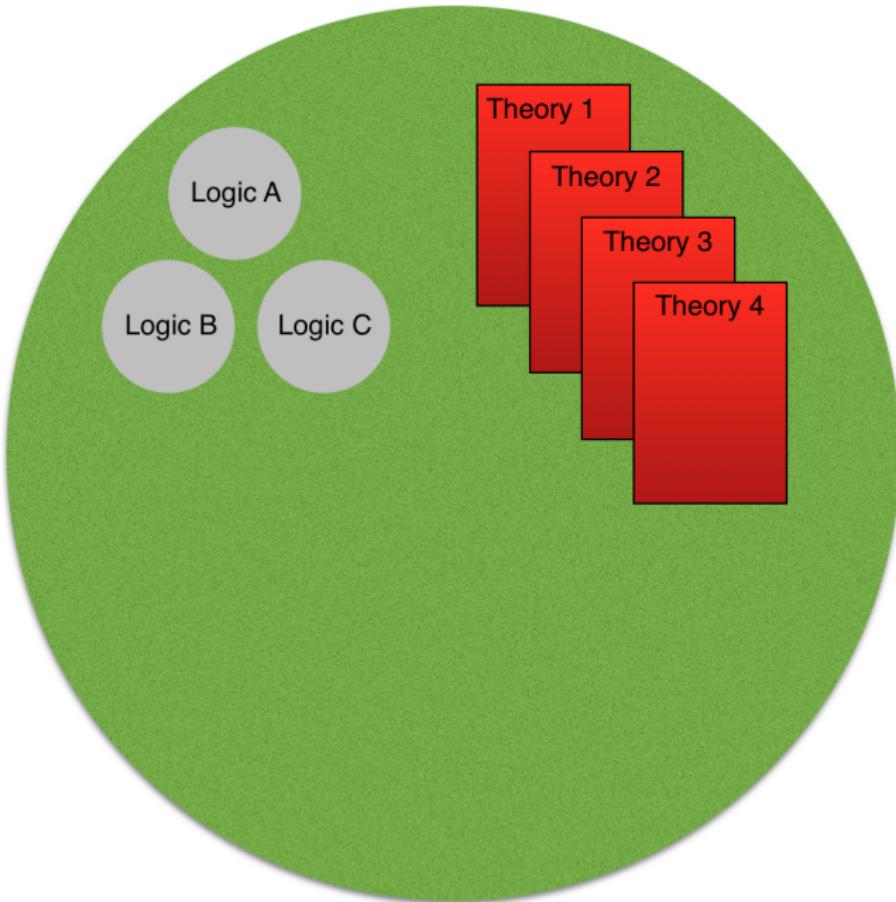
Normative Reasoning Experimentation Platform



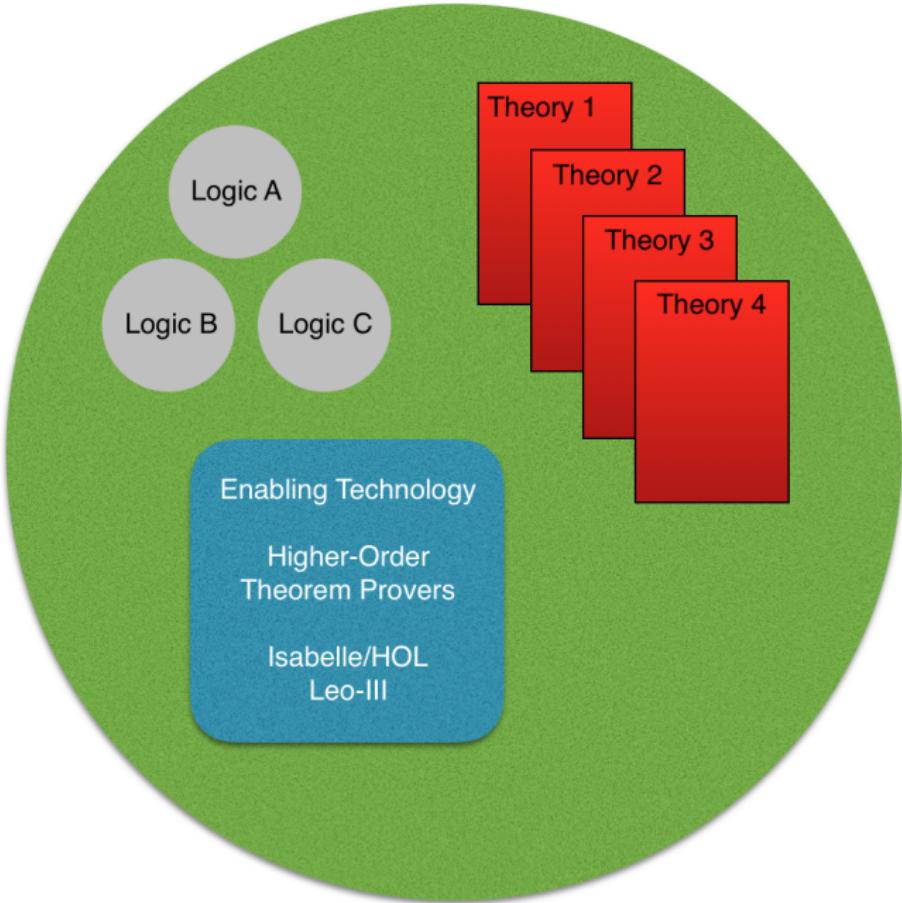
Normative Reasoning Experimentation Platform



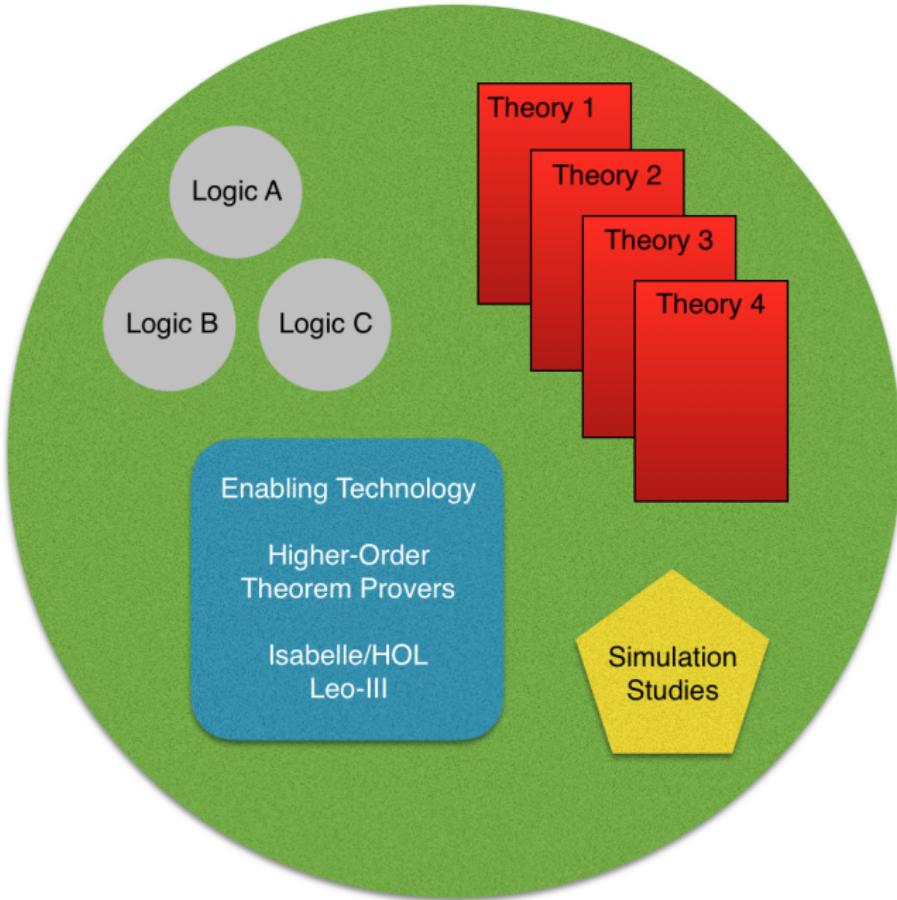
Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform — Demo in Isabelle/HOL

The screenshot shows the Isabelle/HOL interface with the theory file `GDPR.thy` open. The code defines obligations for data processing:

```
1 theory GDPR imports SLD (* Christoph Benzmüller & Xavier Parent, 2018 *)
2
3 begin (** GDPR Example **)
4 consts process_data_lawfully::σ erase_data::σ kill_boss::σ
5
6 axiomatization where
7   (* It is an obligation to process data lawfully. *)
8   A1: " $\text{process\_data\_lawfully} \Rightarrow \text{erase\_data}$ " and
9   (* Implicit: It is an obligation to keep the data if it was processed lawfully. *)
10  Implicit: " $\text{process\_data\_lawfully} \rightarrow \neg \text{erase\_data}$ " and
11  (* If data was not processed lawfully, then it is an obligation to erase the data. *)
12  A2: " $\neg \text{process\_data\_lawfully} \rightarrow \text{erase\_data}$ " and
13  (* Given a situation where data is processed unlawfully. *)
14  A3: " $\neg \text{process\_data\_lawfully} \vdash \text{kill\_boss}$ "
15
16 (** Some Experiments **)
17 lemma True nitpick [satisfy] oops (* Consistency-check: Is there a model? *)
18 lemma False sledgehammer oops (* Inconsistency-check: Can Falsum be derived? *)
19
20 lemma " $\text{erase\_data}$ " sledgehammer nitpick oops (* Should the data be erased? *)
21 lemma " $\neg \text{erase\_data}$ " sledgehammer nitpick oops (* Should the data be kept? *)
22 lemma " $\text{kill\_boss}$ " sledgehammer nitpick oops (* Should the boss be killed? *)
23 end
```

The interface includes a toolbar, a vertical navigation bar on the right, and a status bar at the bottom.

Sledgehammering...

Proof found...

"spass": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could be derived by "e": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could be derived by "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

The screenshot shows the Isabelle/HOL interface with the file 'GDPR.thy' open. The code defines a theory 'GDPR' that imports 'SDL'. It includes an obligation to process data lawfully, an implicit obligation to keep data if processed lawfully, and obligations to erase data or kill a boss if data was not processed lawfully. A 'Danger Zone' box highlights the section where data is processed unlawfully, containing lemmas about whether data should be erased or bosses killed. The interface includes tabs for Documentation, Sidekick, State, and Theories, and a status bar at the bottom.

```
1 theory GDPR imports SDL
(* Christoph Benzmueller & Xavier Parent, 2018 *)
2
3 begin (** GDPR Example **)
4 consts process_data_lawfully::σ erase_data::σ kill_boss::σ
5
6 axiomatization where
7 (* It is an obligation to process data lawfully. *)
8 Al: "[0(process_data_lawfully)]" and
9 (* Implicit: It is an obligation to keep the data if it was processed lawfully. *)
Implicit: "[0(process_data_lawfully → ~erase_data)]" and
10 (* If data was not processed lawfully, then it is an obligation to erase the data. *)
A2: "[~process_data_lawfully → 0(erase_data)]"
11 (* Given a situation where data is processed unlawfully. *) and
12 A3: "[~process_data_lawfully]_cvc"
13
14 (*
15 l
16 l
17 l
18 l
19 l
20 l
21 lemma "[0(~erase_data)]" sledgehammer nitpick oops (* Should the data be kept? *)
22 lemma "[0(kill_boss)]" sledgehammer nitpick oops (* Should the boss be killed? *)
23 end
```

Danger Zone:
Paradoxes and Inconsistencies!

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

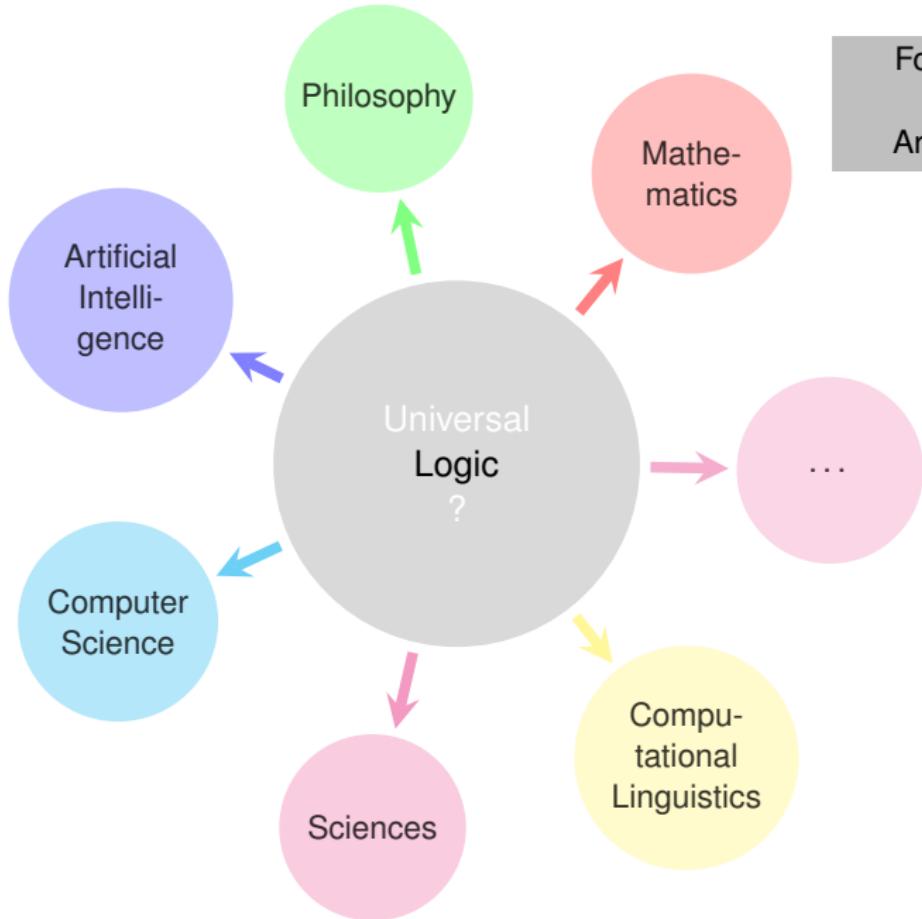
Output Query Sledgehammer Symbols

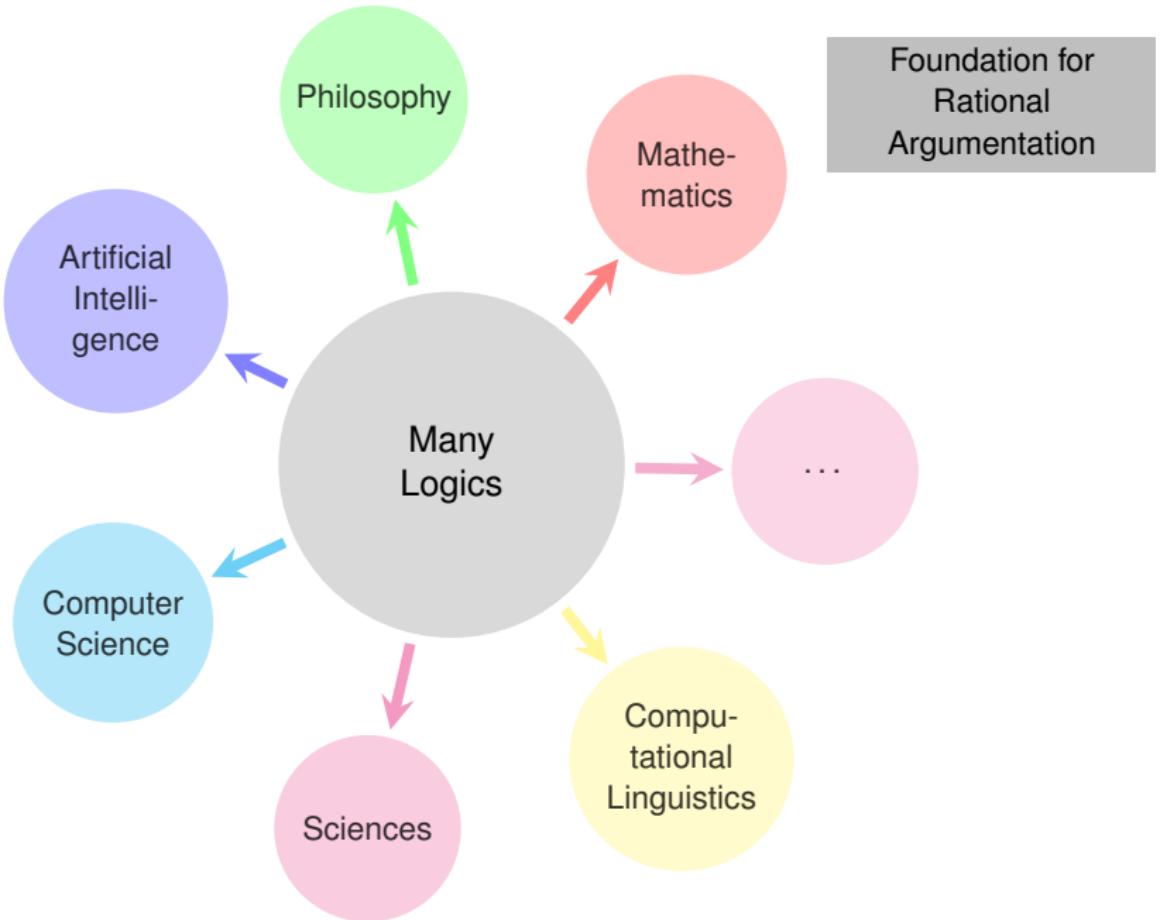
“If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis.”

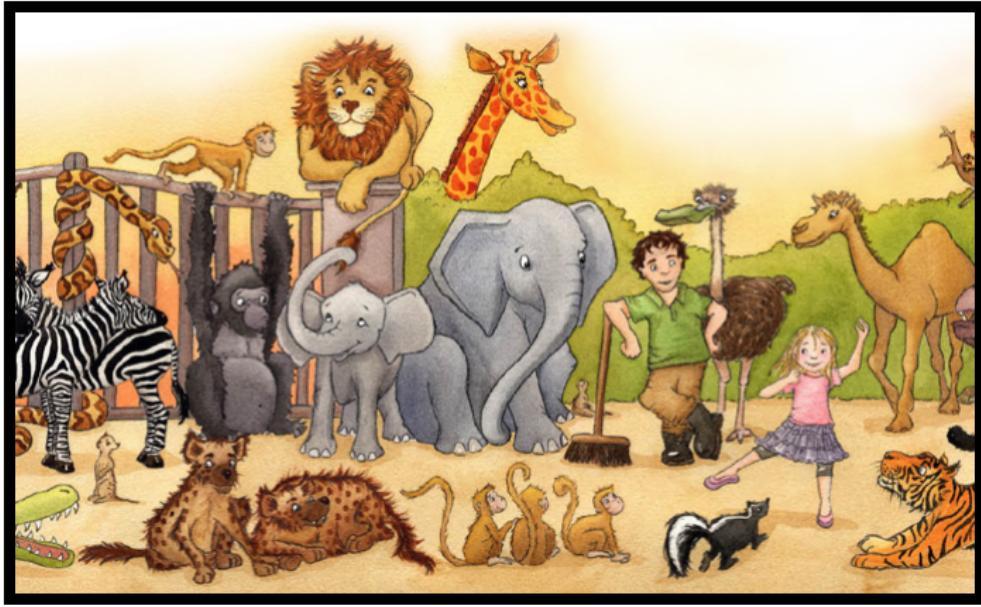
(Leibniz, 1677)

Part B Technology: Universal Reasoning in Higher-Order Logic

Foundation for
Rational
Argumentation







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 Logics, 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:

Kurt Gödel's definition of God:

$$\Box \exists x. Gx$$

$$Gx := \forall \Phi. Positive \Phi \rightarrow \Phi x$$



CONFERENCE

CAMBRIDGE

Formal reconstructive
argument

Günter Eßer - Robert
Gaskins

Received 7 July 2010; Accepted 1 February 2011

© British Society for Analytic Philosophy

Abstract In this paper, we discuss formal reconstructions of Anselm's ontological argument and argue that it contains a formal error. We also show that the proof must therefore fail. The paper shows how such a proof can be easily constructed with the help of a computer system.

Loeben's Ontological Proof of the Existence of God
and the Problem of Divine Simplicity
Wolfgang Löbner (Osnabrück)

Author

The core idea of the ontological proof is to show that the concept of existence is logically prior to all other concepts. Therefore, if one can show that a being that is perfectly actual and exists outside space and time is a logically simple and complete entity (in the sense of Leibniz), then this being is necessarily existent. In this paper, we will show that Loeben's proof of the ontological proof does not contain any errors that would allow us to conclude that the proof is successful. Instead, we will show that there is a serious problem with the proof that makes it impossible to conclude that the proof is successful.

Simpler proofs

No errors in the proof of the "ontological proof"

(2) And a question, does this work?

This paper argues that the proof of the "ontological proof" does not contain any errors that would allow us to conclude that the proof is successful. Instead, we will show that there is a serious problem with the proof that makes it impossible to conclude that the proof is successful.

g (g → g)

THE PROOF IS NOT
FOR THE EXISTENCE OF GOD

LOGICAL ANALYSIS

OF ST. THOMAS ARGUMENTS
JAN SELESNÍK

My main reason for conducting this project was to see whether or not it is possible to construct a proof of the ontological argument that is not based on the concept of God as an infinite being. I have tried to do this with my own framework of the ontological argument, but I have failed to do so.

For a long time now I have accepted the view that the ontological argument is not a good argument for God, but I have never been able to give a good explanation for why it is not a good argument.

Now, however, I have come across a new argument that seems to be able to prove that the ontological argument fails.

I have tried to use it in my own framework, but I have failed to do so. I have, however, been able to give a good explanation for why it is not a good argument. This explanation is based on the fact that the ontological argument fails because it is based on the concept of God as an infinite being. I have not yet been able to give a good explanation for why it is not a good argument. This explanation is based on the fact that the ontological argument fails because it is based on the concept of God as an infinite being. I have not yet been able to give a good explanation for why it is not a good argument. This explanation is based on the fact that the ontological argument fails because it is based on the concept of God as an infinite being.

I have not yet been able to give a good explanation for why it is not a good argument. This explanation is based on the fact that the ontological argument fails because it is based on the concept of God as an infinite being.

A. Kenny (ed.), *God*, Oxford University Press, 1980

APPENDIX A. NOTES IN KURT GÖDEL'S HANDS

[Dated March 19, 1946] The notes in the introduction of Kurt Gödel's *Logic and Theology* are Gödel's, though he does not write the symbols ' Φ ', ' \neg ', and ' \perp '.

(A) $\neg \phi = \neg \psi$ $\neg (\neg \phi \wedge \psi) = \phi \vee \neg \psi$
 $\neg (\phi \vee \psi) = \neg \phi \wedge \neg \psi$ $\neg (\phi \wedge \psi) = \neg \phi \vee \neg \psi$
 $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$ $\neg (\phi \vee \psi) = \neg \phi \wedge \neg \psi$
 $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$ $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$
 $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$ $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$
 $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$ $\neg (\phi \wedge \psi) = \neg \phi \wedge \neg \psi$

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 itself

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?

Magari and others on Gödel's ontological proof

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

1 Introduction

This page is a continuation of my paper [3] and continues about originally published in the third volume of Gödel's collected works [1], but it became available via IOPP when Hajek showed the proof from that book and later pre-

Print Edition
A New Small Entomolof of Gödel's Ontological Proof

Electronic: Ontological proof, Gödel, modal logic, completeness, 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 [1], but it became available via IOPP when Hajek showed the proof from that book and later pre-

Notes
None

Review
The article presents a detailed reconstruction of Gödel's ontological argument and its variants. It discusses the historical development of the argument and shows that it has not been fully understood. It also provides a new perspective on the argument by showing that it can be seen as a form of proof by contradiction. The author argues that Gödel's original proof contains a mistake, which is corrected in the proposed reconstruction. The article also discusses the relationship between Gödel's proof and other proofs of the existence of God, such as Leibniz's and Aquinas'.

Keywords
Ontological proof; Gödel's ontological argument; proof by contradiction; Leibniz; Aquinas

Author's Note
I am grateful to the editor of this journal for his useful comments on this article. I also thank the anonymous referee for his/her useful comments. This research was supported by grants from the Ministry of Education, Culture, and Sports of the Czech Republic (No. 102/07/0144) and from the Grant Agency of the Czech Republic (No. 201/07/0560).

THE ETHICS
(Online: Online Generation, Downloadable)
Translated by R. M. Elmer

PUBLISHER'S CONVENTION:
N.Y.: The Free Press, 1969.

DEFINITIONS:
1. If A is a well-constructed formula of predicate logic, then A is said to be valid if every interpretation of A makes it true.
2. If A is a well-constructed formula of predicate logic, then A is said to be satisfiable if there is an interpretation of A that makes it true.
3. If A is a well-constructed formula of predicate logic, then A is said to be unsatisfiable if there is no interpretation of A that makes it true.

Types, Tableaus, and Gödel's God

by Melvin Fitting

KLUWER ACADEMIC PUBLISHERS

EDWARD N. ZALTA
**ABSTRACT
OBJECTS**
An Introduction to Axiomatic Metatheory

KLUWER PUBLISHING COMPANY
SOFTCOVER: \$40.00 U.S./CANADA

Principia Logico-Metaphysica

(Draft/Excerpt)

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

October 26, 2016

<http://nlp.stanford.edu/principia.pdf>

□(□ A → A) → □ A

THE LOGIC OF PROVABILITY
George Boolos

◆ A → I ◆ A

NORTH-HOLLAND MATHEMATICAL LIBRARY
Categories, Allegories
Peter J. Freyd
Andre Scedrov

North-Holland

Example: Modal Logic Textbook



STUDIES IN LOGIC
AND
PRACTICAL REASONING

VOLUME 3

D.M. GABBAY / P. GARDENFORS / J. SIEKMANN / J. VAN BENTHEM / M. VARDI / J. WOODS

EDITORS

*Handbook of
Modal Logic*

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:

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

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

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

What follows we will 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:

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

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

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):

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

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*,

and V

$V(p)$

$(W, \{$
in the

in a model is a binary relation on W ,
position symbol p in PROP a subset
 p is true. The first two components
model. If there is only one relation
 (W, R, V) for the model itself. We

Metalanguage

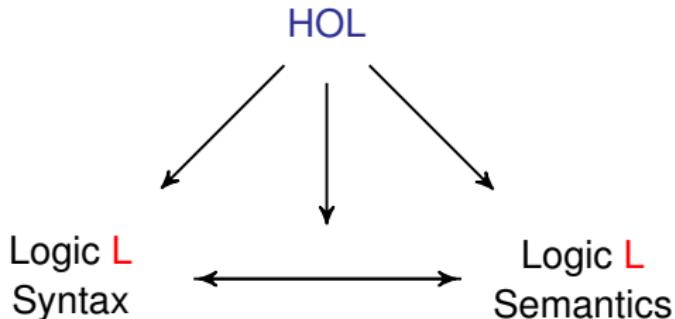
encourage the reader to think of Kripke models as graphs (or to be slightly more precise, *directed graphs*, that is, graphs whose points are linked by directed arrows) and will shortly give some examples which show why this is helpful.

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):

Semantics

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

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, ...

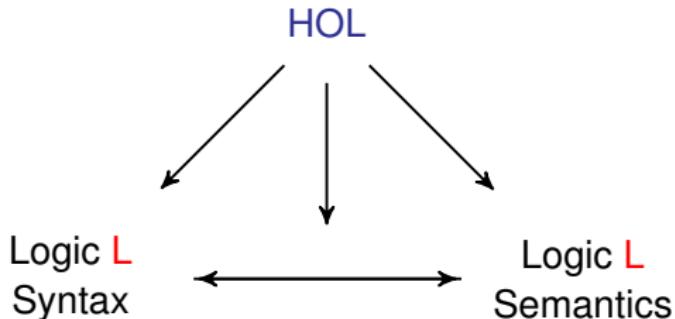
Embedding works also for quantifiers (first-order & higher-order)

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!

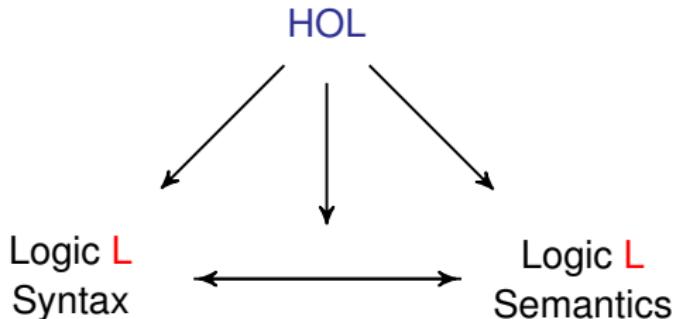
interactive:

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

automated:

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

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, ...

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/Edit 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

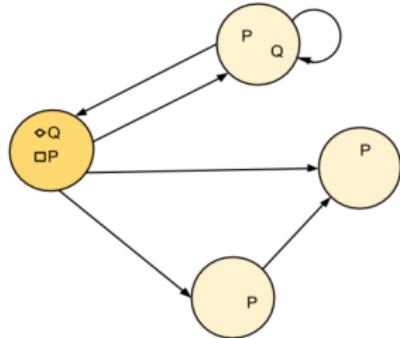
The screenshot shows the Isabelle/HOL IDE interface with the file `GodProof.thy` open. The code defines various modal and quantifier operators using shallow embedding in HOL.

```
1 theory GodProof imports Main
2 begin
3   typedecl i -- "type for possible worlds"
4   typedecl μ -- "type for individuals"
5   type_synonym σ = "(i⇒bool)"
6
7 (* Shallow embedding modal logic connectives in HOL *)
8 abbreviation mneg ("¬_[52]53) where "¬φ ≡ λw. ¬φ(w)"
9 abbreviation mand (infixr "∧" 51) where "φ ∧ ψ ≡ λw. φ(w) ∧ ψ(w)"
10 abbreviation mor (infixr "∨" 50) where "φ ∨ ψ ≡ λw. φ(w) ∨ ψ(w)"
11 abbreviation mimp (infixr "→" 49) where "φ → ψ ≡ λw. φ(w) → ψ(w)"
12 abbreviation mequ (infixr "↔" 48) where "φ ↔ ψ ≡ λw. φ(w) ↔ ψ(w)"
13 abbreviation mnegpred ("¬_[52]53) where "¬Φ ≡ λx. λw. ¬Φ(x)(w)"
14
15 (* Generic box and diamond operators *)
16 abbreviation mboxgen ("□") where "□r φ ≡ λw. ∀v. r w v → φ(v)"
17 abbreviation mdiagon ("◇") where "◇r φ ≡ λw. ∃v. r w v ∧ φ(v)"
18
19 (* Shallow embedding of constant domain quantifiers in HOL *)
20 abbreviation mall_const ("∀c") where "∀c Φ ≡ λw. ∀x. Φ(x)(w)"
21 abbreviation mallB_const (binder "∀c" [8] 9) where "∀c x. φ(x) ≡ ∀c φ"
22 abbreviation mexi_const ("∃c") where "∃c Φ ≡ λw. ∃x. Φ(x)(w)"
23 abbreviation mexiB_const (binder "∃c" [8] 9) where "∃c x. φ(x) ≡ ∃c φ"
24
25 (* Global validity: truth in all possible worlds *)
26 abbreviation mvalid :: "σ ⇒ bool" ("[L][7]110) where "[p] ≡ ∀w. p w"
27
28 (* Shallow embedding of varying domain quantifiers in HOL *)
```

The interface includes a toolbar with icons for file operations, a navigation bar with tabs like Output, Query, Sledgehammer, and Symbols, and a vertical sidebar with tabs for Documentation, Sidekick, State, and Theories.

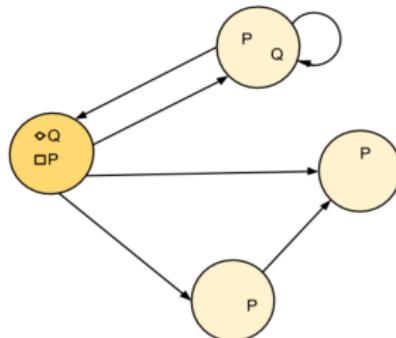
Universal Logic Reasoning in Isabelle/HOL

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



Universal Logic Reasoning in Isabelle/HOL

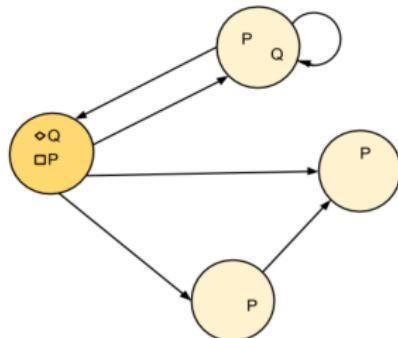
Properties of \Box and \Diamond correlated to structure of transition system between worlds



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

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, $\forall P. \Box P \rightarrow P$
- ▶ Logic KB: symmetric transition relation, $\forall P. P \rightarrow \Box \Diamond P$
- ▶ Logic S5: equivelance relation as transition system, add $\forall P. \Box P \rightarrow \Box \Box P$

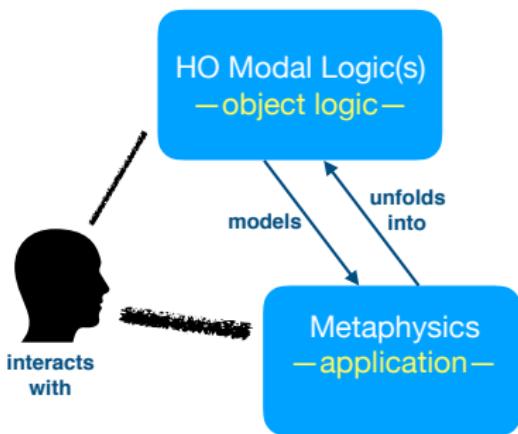
- ▶ Logic D: serial transition relation, $\forall P. \Box P \rightarrow \Diamond P$ (**Standard Deontic Logic**)

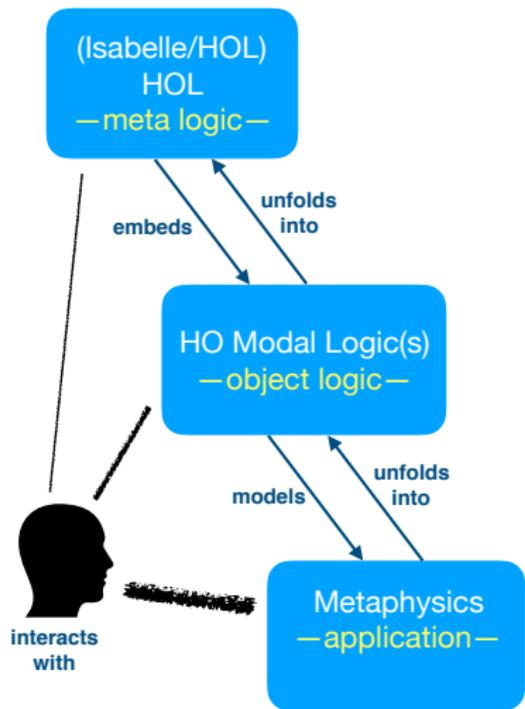


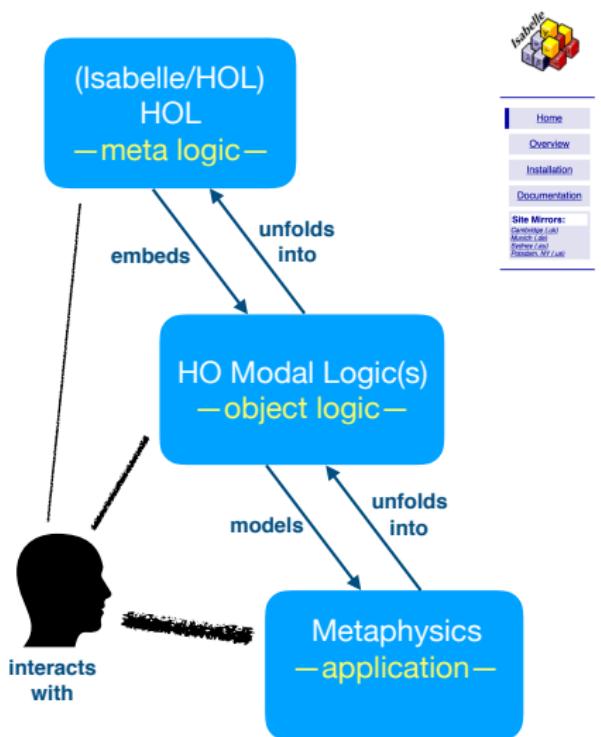
interacts
with



Metaphysics
—application—







[Home](#)
[Overview](#)
[Installation](#)
[Documentation](#)

Site Mirrors:
 Cambridge (UK)
 Munich (Germany)
 Tokyo (Japan)
 Princeton, NJ (USA)

Isabelle

UNIVERSITY OF CAMBRIDGE
 Computer Laboratory
TUM
 Technische Universität München

What is Isabelle?

Isabelle is a generic proof assistant. It allows mathematical formulae to be expressed in a formal language and provides tools for proving those formulae 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 Linux](#) - [Download for Windows \(32bit\)](#) - [Download for Windows \(64bit\)](#) - [Download for Mac OS X](#)

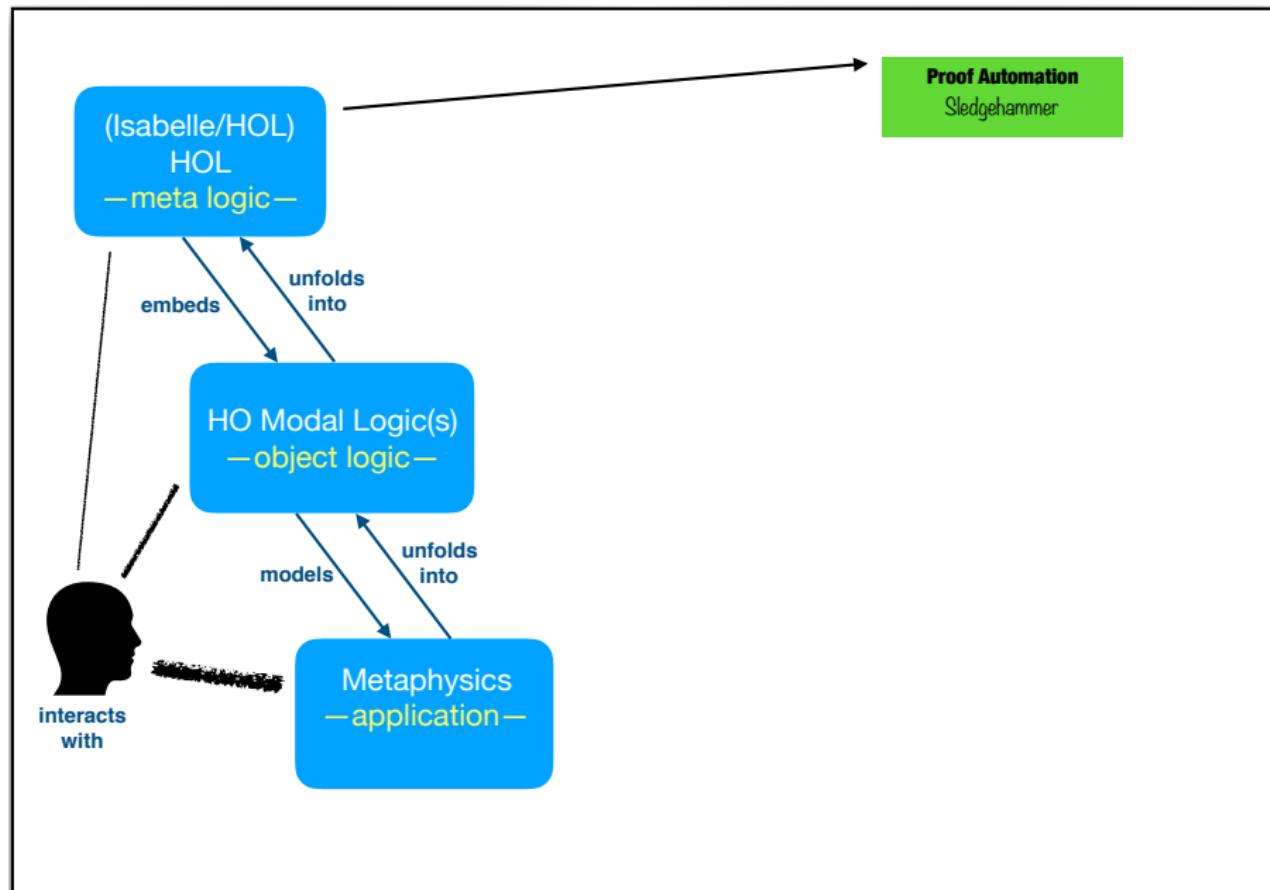
Some notable changes:

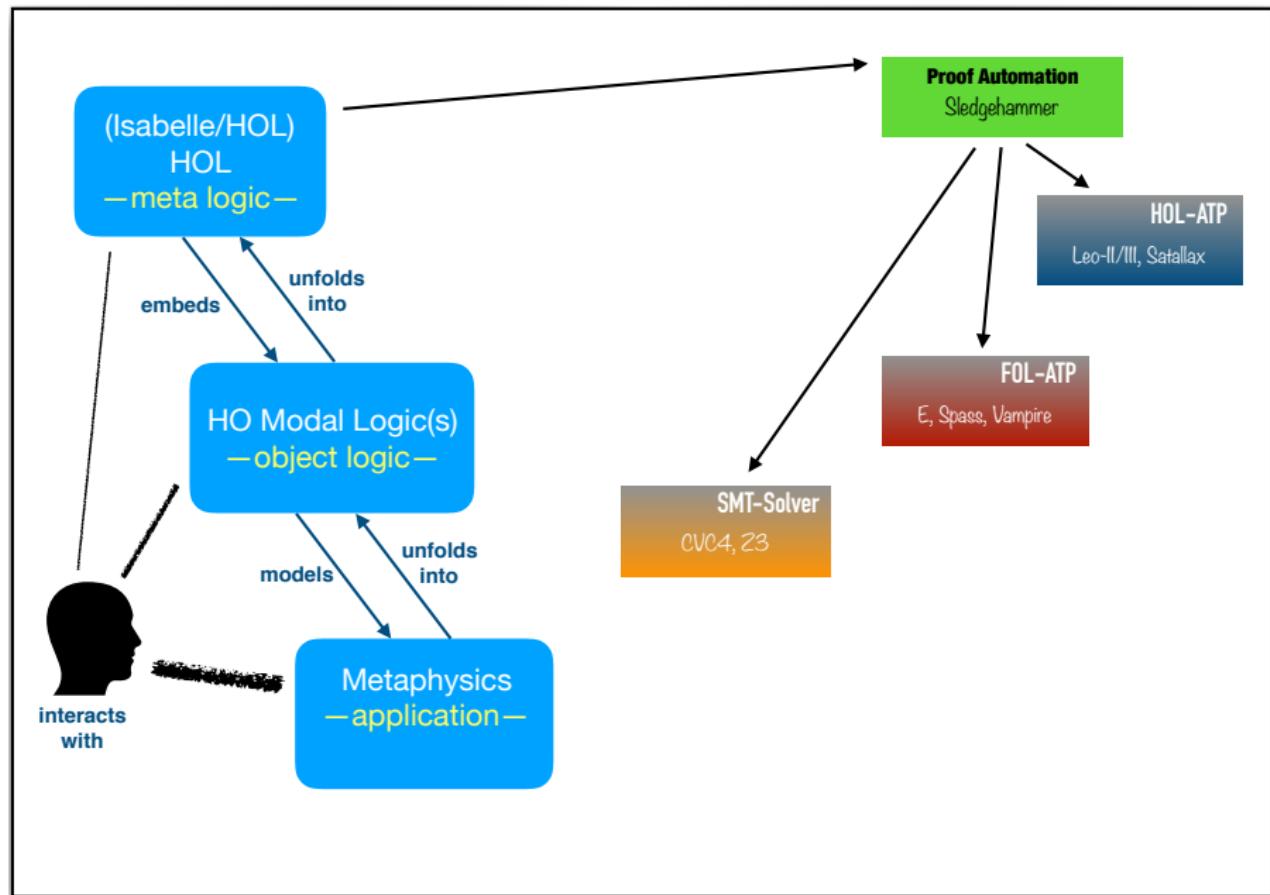
- Experimental support for Visual Studio Code as alternative PIDE front-end.
- Improved Isabelle/Edits Prover IDE: management of session sources independently of editor buffers, removal of unused theories, explicit indication of theory status, more careful auto-indentation.
- Session-qualified theory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

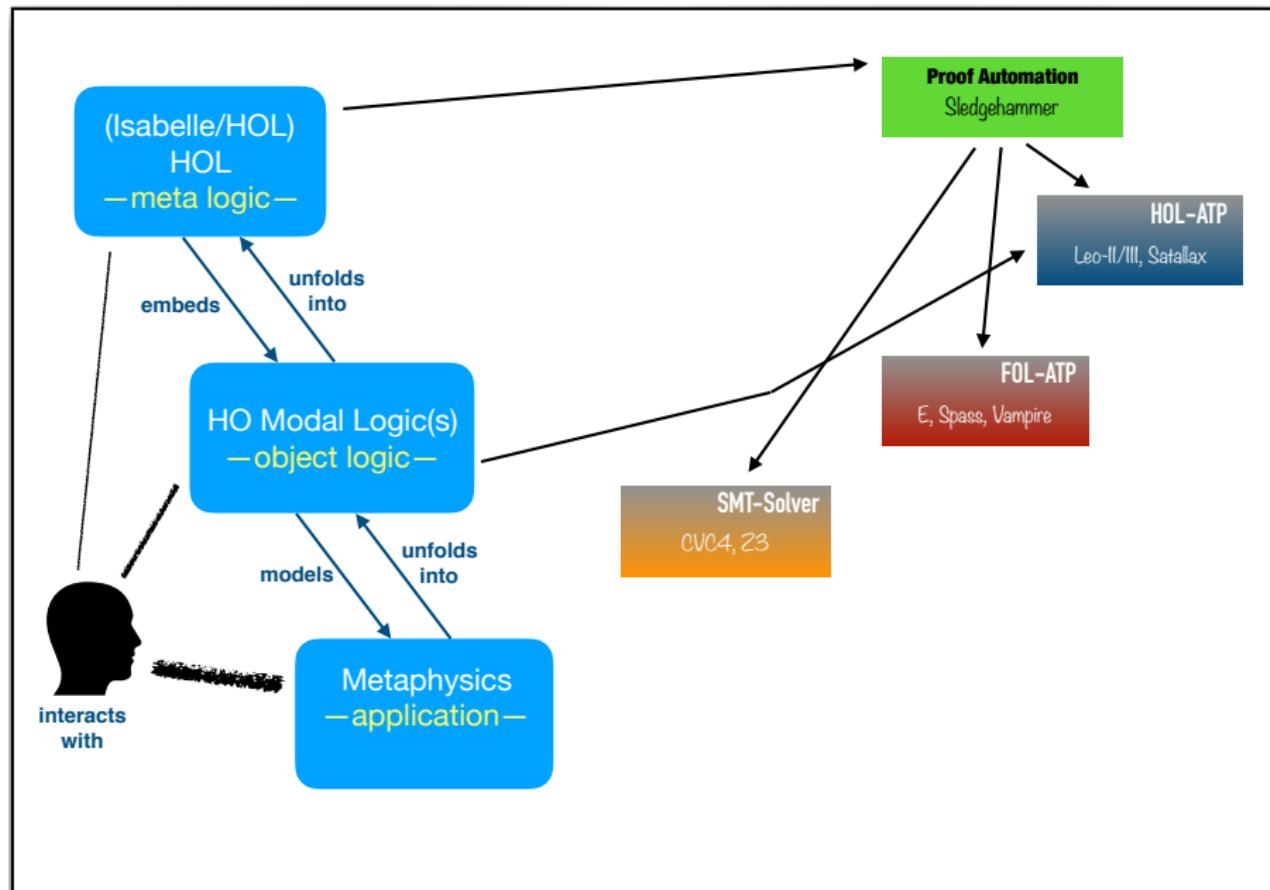
See also the cumulative [NEWS](#).

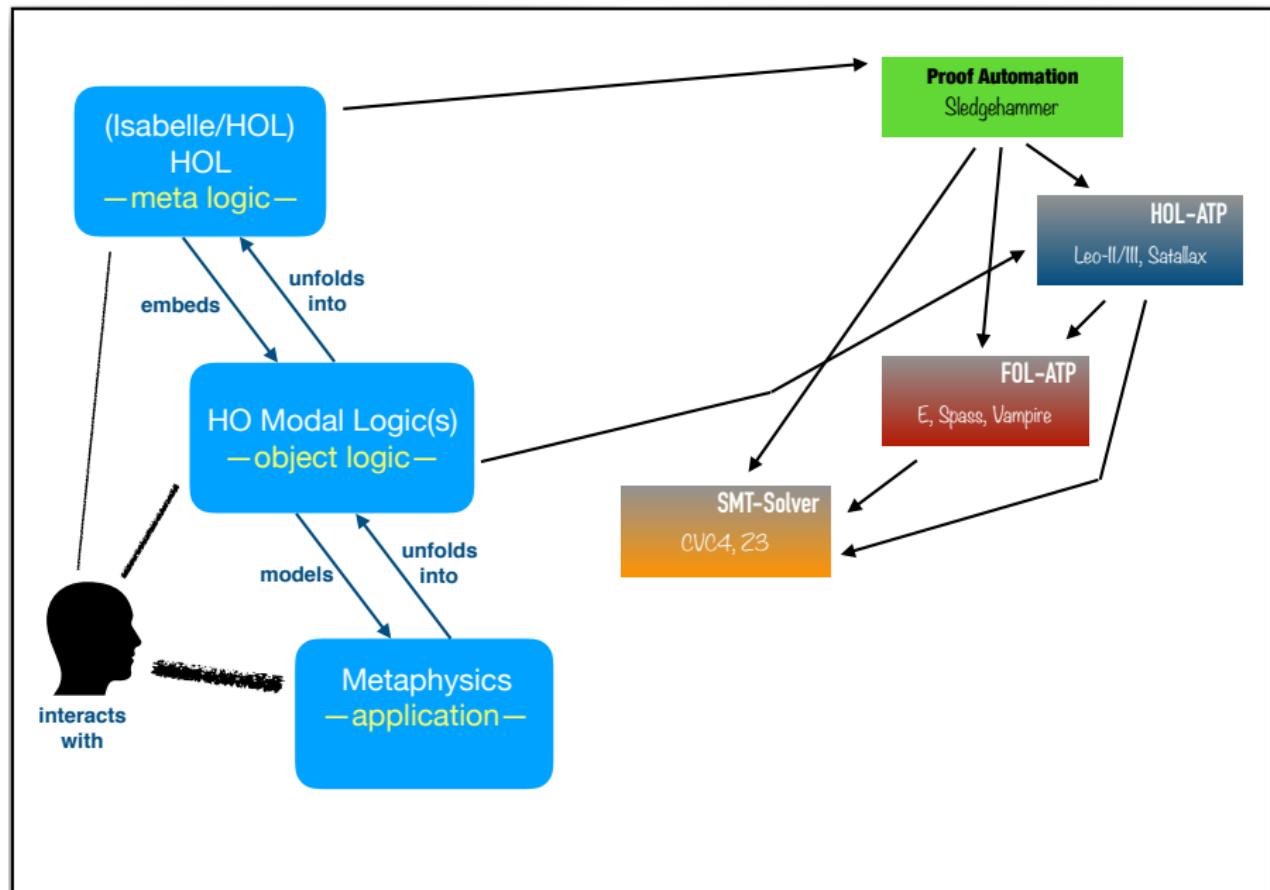
Distribution & Support

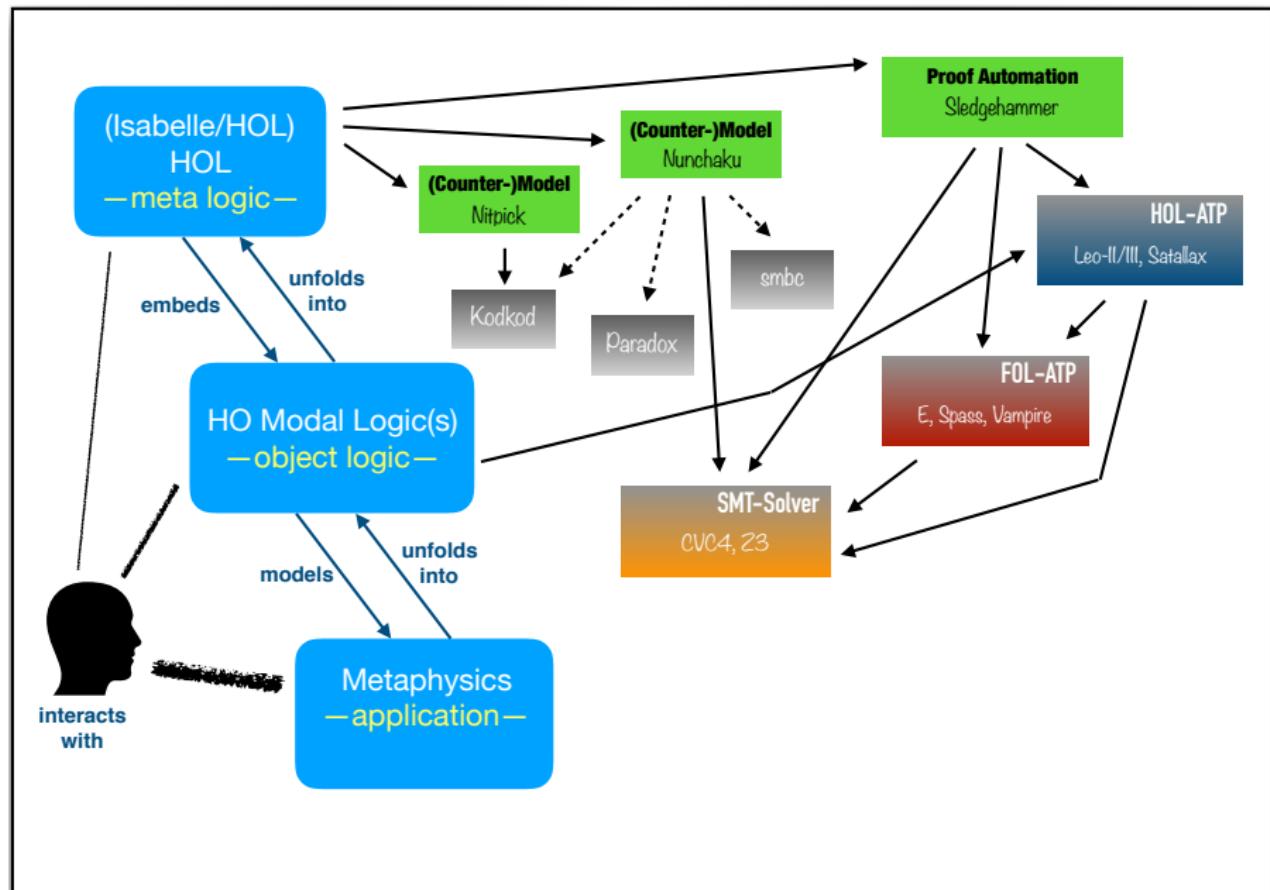
Isabelle is distributed for free under a conglomeration of open-source licenses, but the main code-base is subject to BSD-style regulations. The application bundles include source and binary packages and documentation, see the detailed [installation instructions](#). A vast collection of Isabelle examples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

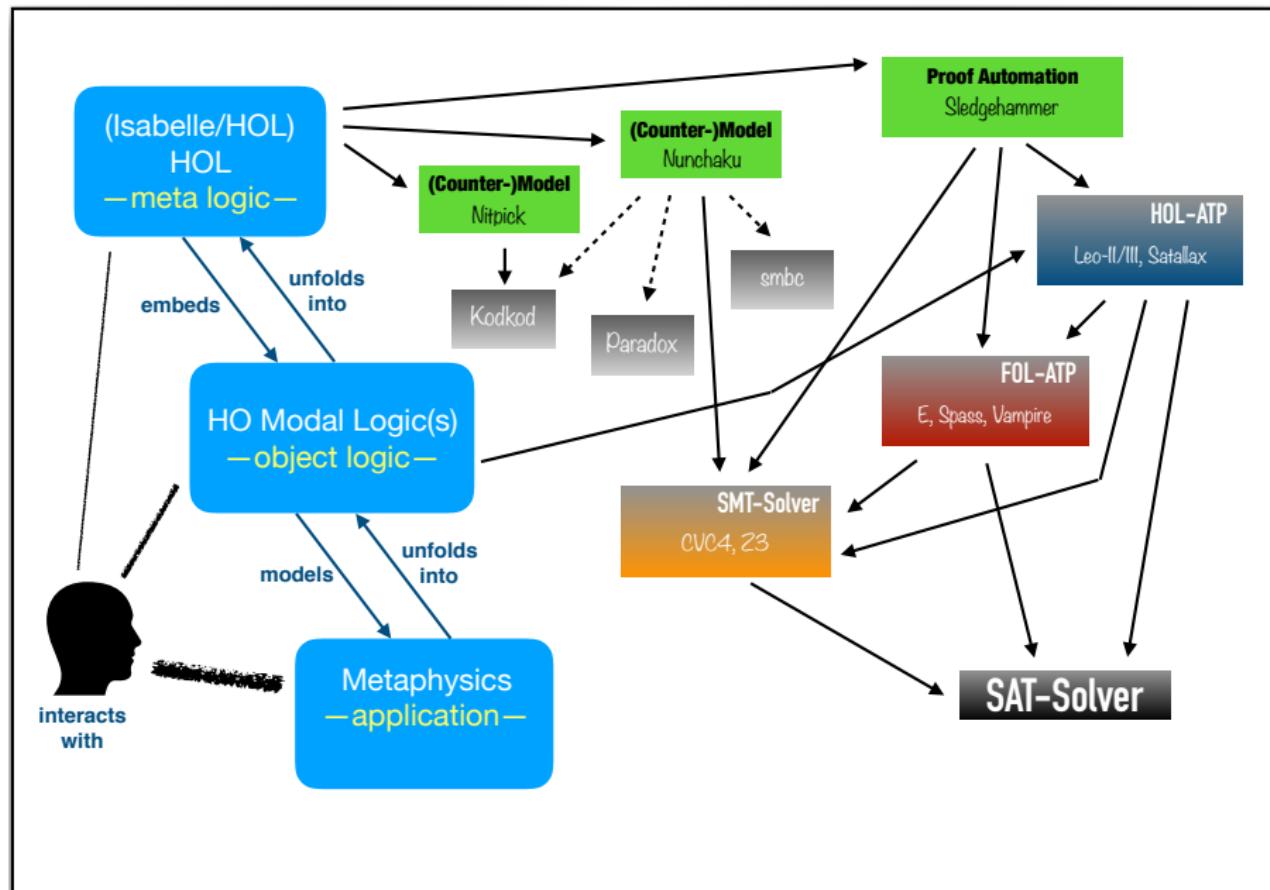


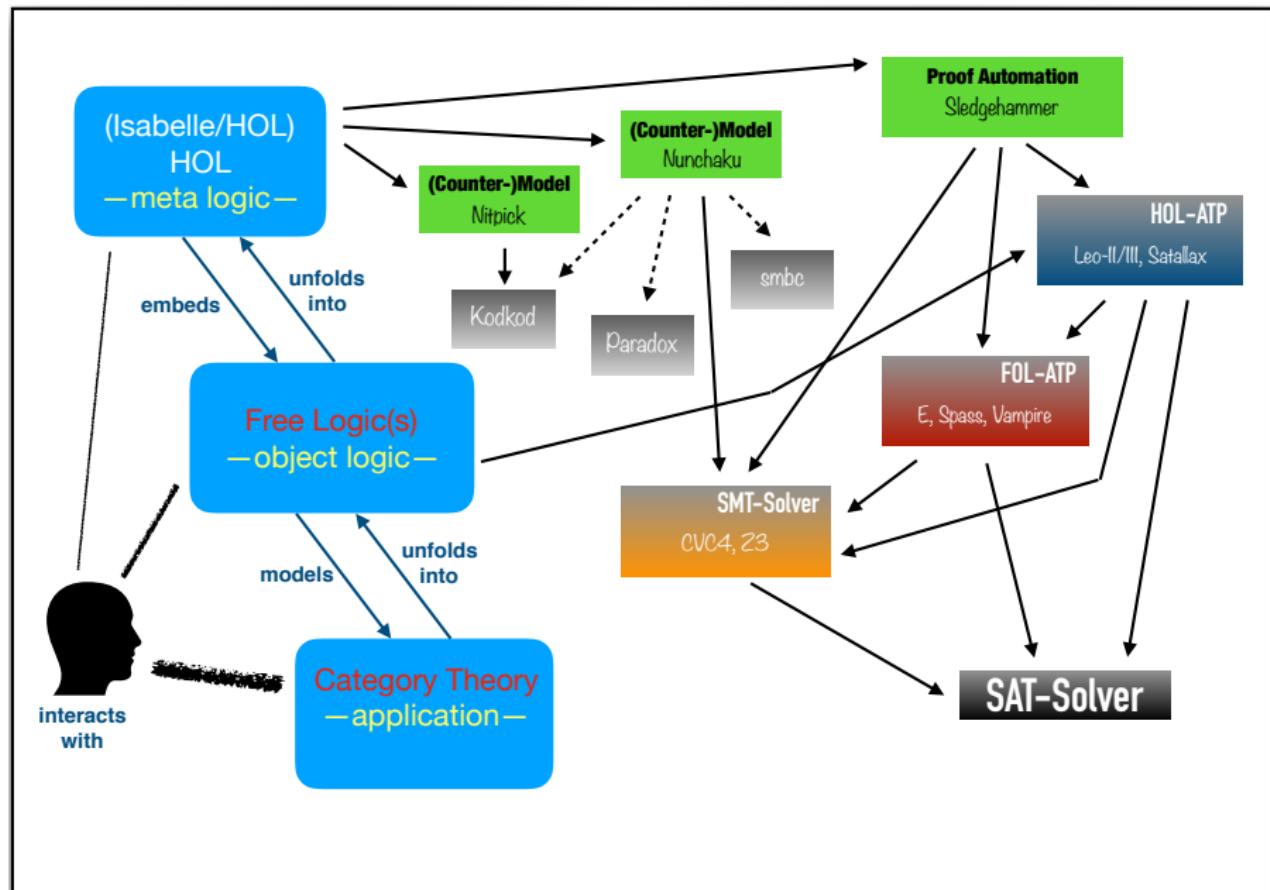


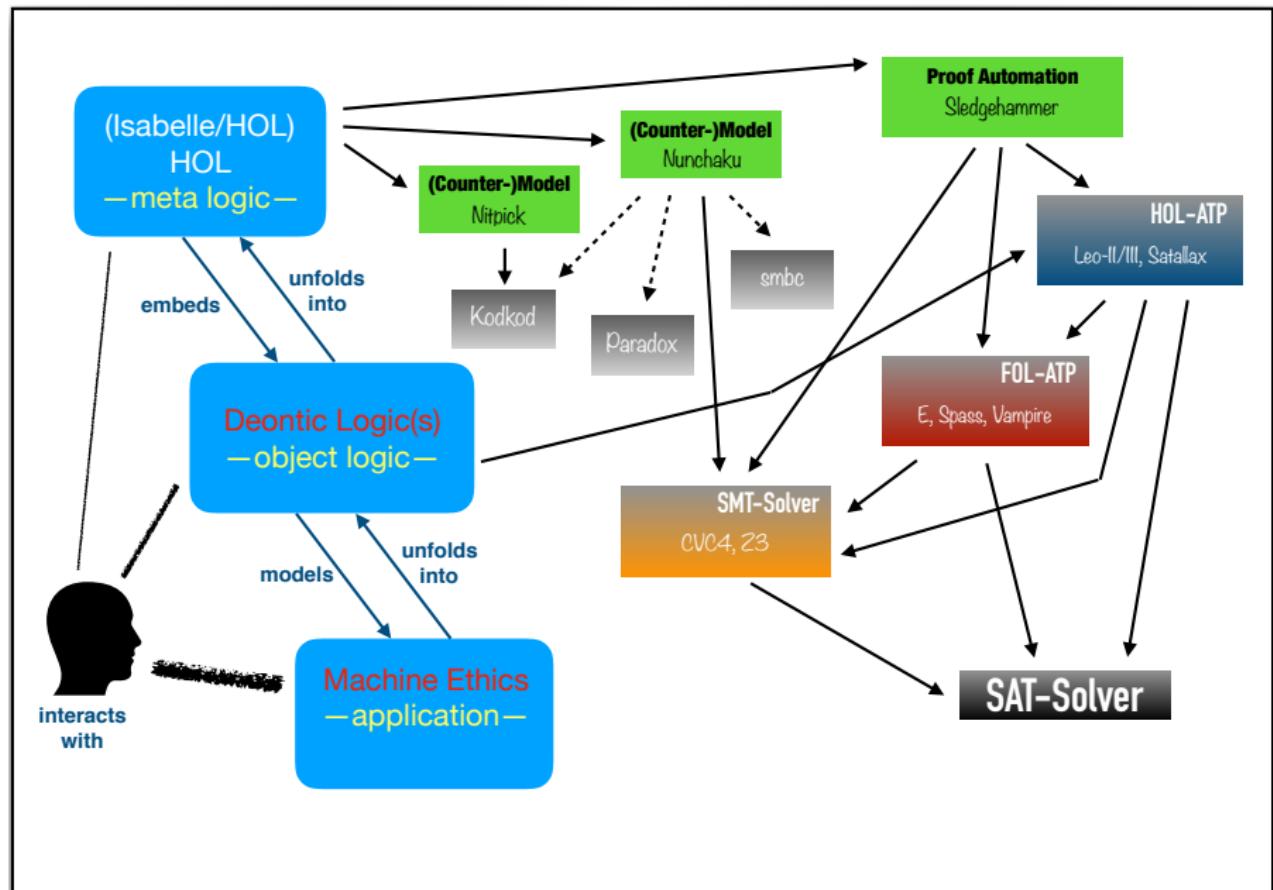


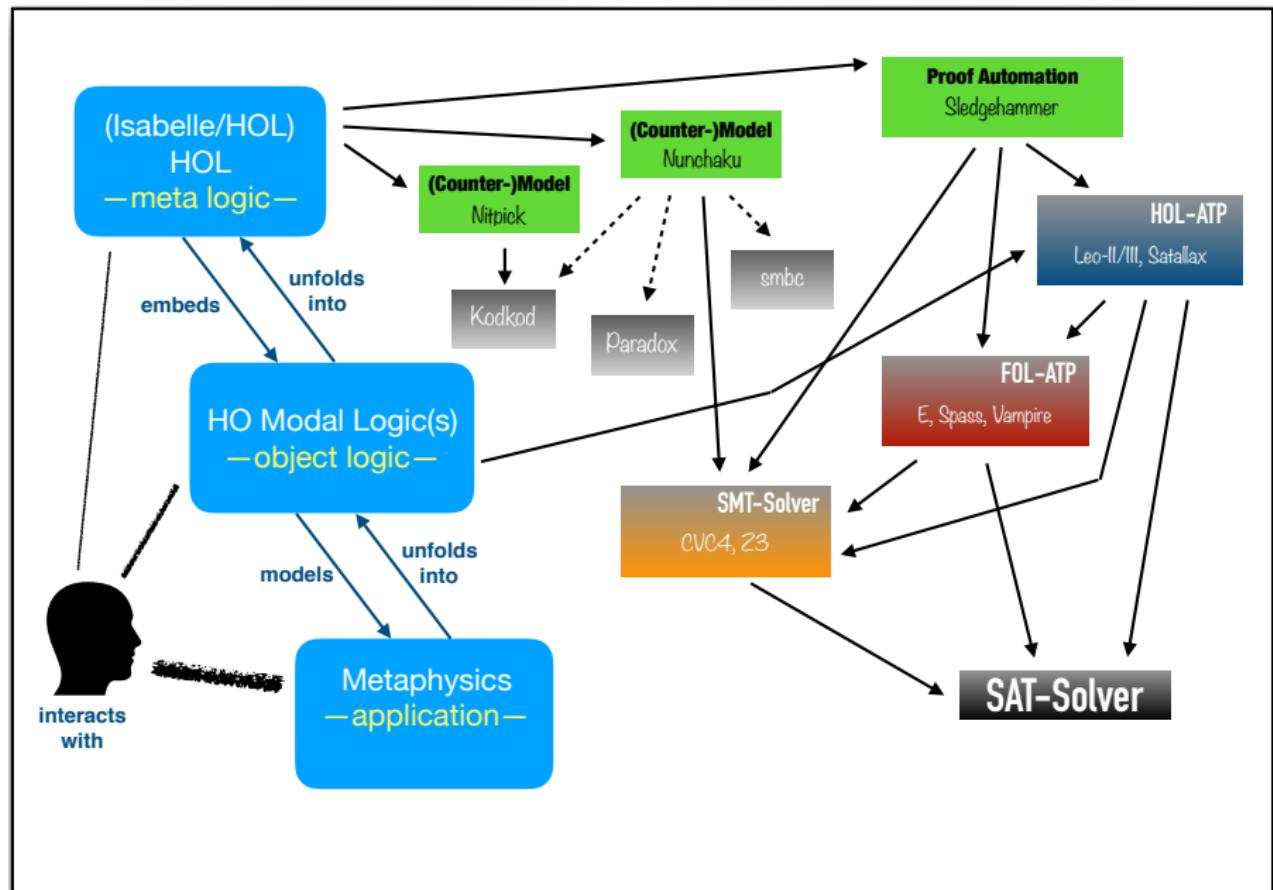


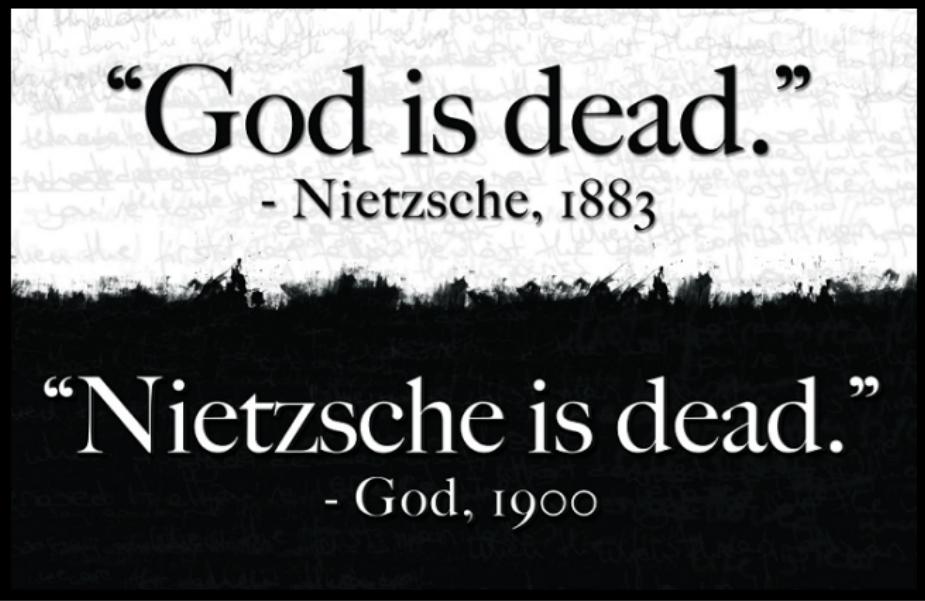












“God is dead.”

- Nietzsche, 1883

“Nietzsche is dead.”

- God, 1900

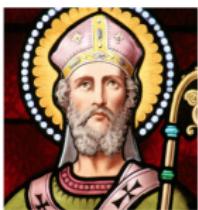
Part C

Evidence: Analysing Rational Arguments in Metaphysics

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

Ontological Proofs of God's Existence

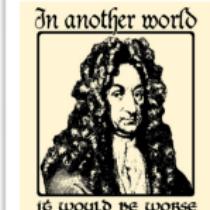
A Long and Continuing Tradition in Philosophy



St. Anselm



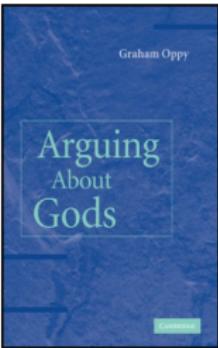
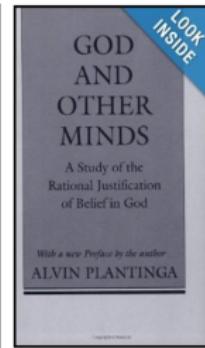
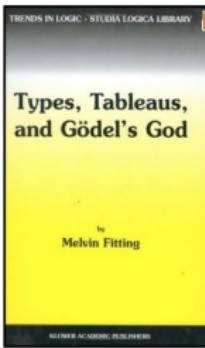
Descartes



Leibniz



Gödel



Computational Metaphysics: Kurt Gödel's Ontological Argument

Ontologischer Beweis Feb 10, 1970

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

At 1 $P(\varphi), P(\psi) \supset P(\varphi \wedge \psi)$ At 2 $P(\varphi) \supset P(\neg \varphi)$

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

P2 $\varphi \text{ Emx} \equiv (\psi)[\forall x(\psi(x) \supset N(y)[\varphi(y) \supset \psi(y)])]$ (Existence)

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

At 2 $\begin{array}{l} P(\varphi) \supset N P(\varphi) \\ \sim P(\varphi) \supset N \sim P(\varphi) \end{array} \left. \begin{array}{l} \text{because it follows} \\ \text{from the nature of the} \\ \text{property} \end{array} \right\}$

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

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

At 3 $P(E)$

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

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

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

 " $\supset N(\exists y) G(y)$ M = permuting

any two instances of x are met. equivalent
exclusive or and for any number of numerants

$M(\exists x) G(x)$ means "all
possible This is:
At 4: $P(\varphi) \cdot \varphi \supset \varphi$

~~the system~~
 ~~$\{x=x$ is pr~~
 ~~$\{x \neq x$ is pr~~

But if a system is y ,
it would mean, that the num.prop. is (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 ~~the~~ ⁱⁿ the at. time. It me-
also mean "affirmation" as opposed to "privatism"
(or catastrophic privation). This supports the pl. part

$\neg \varphi$ is negative $\neg (\varphi \supset \varphi)$ Otherwise $\varphi(x) \supset x \neq x$
hence $x \neq x$ positive $\neg x = x$ negative At
or the epistem. part

i.e. the formal 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 ($\Leftrightarrow \varphi \in P$)

At 1 $P(\varphi), P(\psi) \supset P(\varphi \wedge \psi)$ • At 2 $P(\varphi) \supset P(\neg \varphi)$

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

P2 $\varphi \text{ Emx} \equiv (\psi)[\forall x(\psi(x) \supset N(y)[\varphi(y) \supset \psi(y)])]$ (Existence)

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

At 2 $\begin{array}{l} P(\varphi) \supset N(P(\varphi)) \\ \sim P(\varphi) \supset N \sim P(\varphi) \end{array} \left. \begin{array}{l} \text{because it follows} \\ \text{from the nature of the} \\ \text{property} \end{array} \right\}$

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

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

At 3 $P(E)$

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

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

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

" $\supset N(\exists y) G(y)$ $M = \text{functionality}$

any two instances of x are nec. equivalent

exclusive or * and for any number of numerants

$M(x) G(x)$: means all pos. prop. w.r.t. com-
patible
This is true because of:

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

~~True~~ { $x=x$ is positive
~~False~~ { $x \neq x$ is negative

But if a system S of pos. prop. were incon-
sistent it would mean, that the num.prop. x (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 ~~in the at. time~~ It may
also mean "affirmation" as opposed to "privat-
(or ~~containing~~ negation). This supports the pl. part

\supset of φ positive $\exists x N \sim P(x)$ Otherwise $\varphi(x) \supset x \neq x$
hence $x \neq x$ positive $\exists x x \neq x$ negating At.
or the exist. of pos. At.

~~dog~~
i.e. the formal 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:

$$\Diamond\exists xG(x)$$

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: $\phi \text{ ess. } x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being:

$$\forall x[G(x) \rightarrow G \text{ ess. } x]$$

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

Axiom A5 Necessary existence is a positive property:

$$P(NE)$$

Thm. T3 Necessarily, God exists:



$$\Box\exists xG(x)$$

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:

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

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:

$$\Diamond\exists xG(x)$$

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: $\phi \text{ ess. } x \leftrightarrow \phi(x) \wedge \Box\forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

Thm. T2 Being God-like is an essence of any God-like being:

$$\forall x[G(x) \rightarrow G \text{ ess. } x]$$

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

Axiom A5 Necessary existence is a positive property:

$$P(NE)$$

Thm. T3 Necessarily, God exists:

$$\Box\exists xG(x)$$

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:

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

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:

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

Thm. T2 Being God-like is an essence of any God-like being:

$$\forall x [G(x) \rightarrow G \text{ ess. } x]$$

Def. D3 Necessary existence of an individual is the necessary exemplification of all its essences:

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

Axiom A5 Necessary existence is a positive property:

$$P(NE)$$

Thm. T3 Necessarily, God exists:

$$\Box \exists x G(x)$$

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:

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

Axiom A3 The property of being God-like is positive:

$$P(G)$$

Cor. C Possibly, God exists:

$$\Diamond\exists xG(x)$$

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:

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

Thm. T2 Being God-like is an essence of any God-like being:

$$\forall x[G(x) \rightarrow G \text{ ess. } x]$$

Def. D3 Necessary existence of an individual is the necessary exemplification of all its essences:

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

Axiom A5 Necessary existence is a positive property.

$$P(NE)$$

Thm. T3 Necessarily, God exists:

$$\Box\exists xG(x)$$

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: $\Diamond\exists xG(x)$

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: $\phi\text{ ess. }x \leftrightarrow \phi(x) \wedge \forall\psi(\psi(x) \rightarrow \Box\forall y(\phi(y) \rightarrow \psi(y)))$

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

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

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

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

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

$$\Diamond\exists xG(x)$$

Axiom A4

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

Def. D2

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

Thm. T2

$$\forall x[G(x) \rightarrow G \text{ ess. } x]$$

Def. D3

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

Axiom A5

$$P(NE)$$

Thm. T3

$$\Box\exists xG(x)$$

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

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

Def. D3

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

Axiom A5

$$P(NE)$$

Thm. T3

$$\Box\exists xG(x)$$

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

The screenshot shows the GodProof theorem prover interface. The main window displays a file named "GodProof.thy" containing a proof script. The script includes definitions for "ess" (essence), "T2" (Being God-like), "NE" (Necessary Existence), and "A5" (Necessary Existence is a positive property). It also contains a theorem "T3" (Necessarily, God exists) proved using the "sledgehammer" tactic. The interface features a toolbar at the top, a vertical sidebar on the right with tabs for Documentation, Sidekick, State, and Theories, and a status bar at the bottom.

```
124 (* Ess: An essence of an individual is a property possessed by
125     it and necessarily implying any of its properties: *)
126 definition ess (infixr "ess" 85) where
127   " $\Phi \text{ ess } x = \Phi x \wedge (\forall^{\exists} \Psi. \Psi(x) \rightarrow \Box(\forall^{\exists} y. \Phi(y) \rightarrow \Psi(y)))$ "
128
129 (* T2: Being God-like is an essence of any God-like being *)
130 theorem T2: " $[\forall^{\exists} x. G(x) \rightarrow G \text{ ess } x]$ " by (metis A1b A4 G_def ess_def)
131
132 (* NE: Necessary existence of an individual is the necessary
133    exemplification of all its essences *)
134 definition NE where " $NE(x) = (\forall^{\exists} \Phi. \Phi \text{ ess } x \rightarrow \Box(\exists^{\exists} y. \Phi(y)))$ "
135
136 (* A5: Necessary existence is a positive property *)
137 axiomatization where A5: "[P(NE)]"
138
139 (* T3: Necessarily, God exists *)
140 theorem T3: " $[\Box(\exists^{\exists} x. G(x))]$ "  
sledgehammer  
sledgehammer [remote_leo2 remote_satallax]  
by (metis A5 C G_def NE_def KB T2)
```

Sledgehammering...
Proof found...
"remote_satallax": Timed out
"remote_leo2": Try this: by (metis A1a A1b A2 A3 A4 A5 C NE_def S4 S5 T1 T2 ess_c

Output Query Sledgehammer Symbols

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 justified
- ▶ 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 justified
- ▶ 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 justified
- ▶ 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 justified
- ▶ 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 justified
- ▶ 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.

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: $\varphi \rightarrow \Box \varphi$

- ▶ 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: $\varphi \rightarrow \Box \varphi$

- ▶ 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: $\varphi \rightarrow \Box \varphi$



- ▶ 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.

Petr Hájek

A New Small Emendation of Gödel's Ontological Proof

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

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

Er ist gut, daß wir nicht wissen,
sondern glauben, daß er Gott sei.
(Kant, Nachleß)

1. Einführung

Gödels zu Lebzeiten unveröffentlichter Beweis für die notwendige Existenz eines Gott-ähnlichen Wesen hat sowohl philosophisches als auch mathematisches Interesse geweckt. In einer späteren, unveröffentlichten Arbeit ist es, zu einer Deutung dieses Beweises gekommen, die durch Kommentierung des Schriftgegenstands „Gödel's Ontological Proof“ durch Bereitstellung von etwas Modelltheorie, die Arbeit mit dem Lehrsatz philosophisches Beimengen. Während der letzten Jahre habe ich etliche Male über Gödels Gottesbeweis vorgetragen, insbesondere auf dem Symposium zur Feier von Professor Gert Müller (Heidelberg, Januar 1991), doch habe ich nievals beobachtigt, eine Veröffentlichung über das Thema zu machen. Da ich wiederholt um eine schriftliche Version gebeten wurde, einschließlich mich, schnell eine „erweiterte Kürzfassung“¹ zu schreiben, ohne aus ihr einen

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@iuivt.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

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.

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

Es ist gut, daß wir nicht wissen,
sondern glauben, daß es Gott sei.
(Karl Maxfeld)

1. Einführung

Gödels zu Lebzeiten unveröffentlichter Beweis für die notwendige Existenz eines Gott-ähnlichen Wesen hat sowohl philosophisches als auch mathematisches Interesse geweckt. Zwei durch Kommentatoren entdeckte Varianten dieses Beweises sind durch Bereitstellung von etwas Modelltheorie, insbesondere einer logisch-philosophischen Beiträge. Während der letzten Jahre habe ich etliche Male über Gödels Gottesbeweis vorgetragen, insbesondere auf dem Symposium zur Feier von Professor Gert Müller (Heidelberg, Januar 1991), doch habe ich nievals beachtigt, eine Veröffentlichung über das Thema zu machen. Da ich wiederholt um eine schriftliche Version gebeten wurde, einschließlich mich, schnell eine „erweiterte Kürzfassung“¹ zu schreiben, ohne aus ihr einen

Gödel's Ontological Proof Revisited *

C. Anthony Anderson and Michael Gettings

University of Cambridge, United Kingdom
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.

Petr Hájek

A New Small Emendation of Gödel's Ontological Proof

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

1. Introduction

Gödel's ontological proof of necessary existence of a godlike being was finally published in his 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) in a seminar at Princeton. Detailed history is found in Adams' introduction 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

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@iuivt.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

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

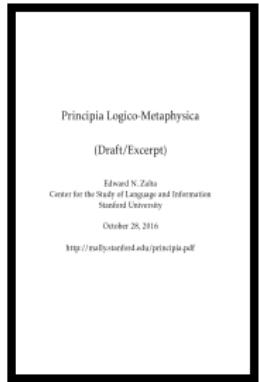
Hyperintensional higher-order modal logic

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



Ed Zalta (Stanford)

Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected



Daniel Kirchner
(Mathematics, FU Berlin)

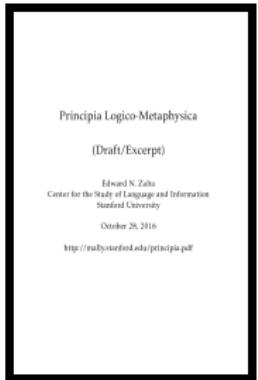
Kirchner Paradox

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



Ed Zalta (Stanford)

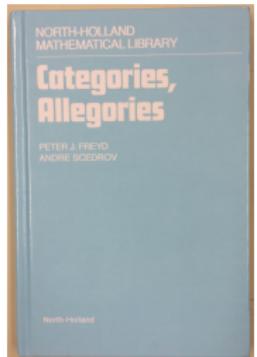
Principia Logico-Metaphysica

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



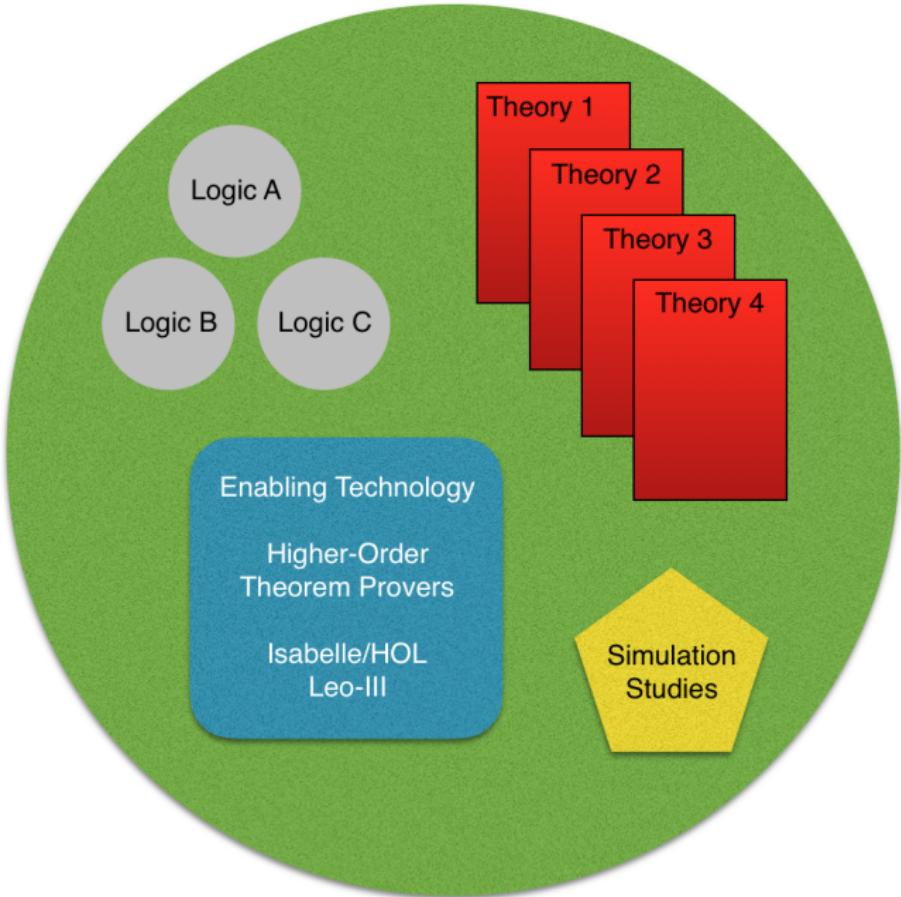
D. Scott
(UC Berkeley)

“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



The screenshot shows the Isabelle/HOL interface with the theory file 'SDL.thy' open. The code defines various modal operators and properties of the logic.

```

1 theory SDL imports Main          (* Christoph Benzmueller & Xavier Parent, 2018 *)
2
3 begin (* SDL: Standard Deontic Logic (Modal Logic D) *)
4 typecl型 i (*type for possible worlds*) type_synonym σ = "(i⇒bool)"
5 consts r::"i⇒i⇒bool" (infixr "r" 70) (*Accessibility relation.*) cw::i (*Current world.*)
6
7 abbreviation mtop ("T") where "T ≡ λw. True"
8 abbreviation mbot ("⊥") where "⊥ ≡ λw. False"
9 abbreviation mnnot ("¬"[52]53) where "¬φ ≡ λw. ¬φ(w)"
10 abbreviation mand (infixr "∧" 51) where "φ∧ψ ≡ λw. φ(w) ∧ ψ(w)"
11 abbreviation mor (infixr "∨" 50) where "φ∨ψ ≡ λw. φ(w) ∨ ψ(w)"
12 abbreviation mimp (infixr "→" 49) where "φ→ψ ≡ λw. φ(w) → ψ(w)"
13 abbreviation mequ (infixr "↔" 48) where "φ↔ψ ≡ λw. φ(w) ↔ ψ(w)"
14 abbreviation mobligatory ("OB") where "OB φ ≡ λw. ∀v. w r v → φ(v)" (*obligatory*)
15 abbreviation mpermissible ("PE") where "PE φ ≡ ¬(OB(¬φ))" (*permissible*)
16 abbreviation impermissible ("IM") where "IM φ ≡ OB(¬φ)" (*impermissible*)
17 abbreviation omissible ("OM") where "OM φ ≡ ¬(OB φ)" (*omissible*)
18 abbreviation moptional ("OP") where "OP φ ≡ (¬(OB φ) ∧ ¬(OB(¬φ)))" (*optional*)
19
20 abbreviation ddlderived::"σ ⇒ bool" ("[_]"[7]105)      (*Global Validity*)
21   where "[A] ≡ ∀w. A w"
22 abbreviation ddlderivedcw::"σ ⇒ bool" ("[_]cw"[7]105) (*Local Validity (in cw)*)
23   where "[A]cw ≡ A cw"
24
25 (* The D axiom is postulated *)
26 axiomatization where D: "[¬ ((OB φ) ∧ (OB (¬ φ)))]"
27
28 (* Meta-level study: D corresponds to seriality *)
29 lemma "[¬ ((OB φ) ∧ (OB (¬ φ)))] ← (λw. ∃v. w r v)" by auto
30
31 (* Standardised syntax: unary operator for obligation in SDL *)
32 abbreviation obligatorSDL::"σ⇒σ" ("0[_]") where "0(A) ≡ OB A"
33
34 (* Consistency *)
35 lemma True nitpick [satisfy] oops

```

The right sidebar shows navigation links: Documentation, Sidekick, State, Theories.

Bottom navigation bar: Output, Query, Sledgehammer, Symbols.

Demo I: DDL in Isabelle/HOL

DDL.thy

```

1 theory DDL imports Main (* Christoph Benzmueller & Xavier Parent & Ali Farjami, 2018 *)
2
3 begin (* DDL: Dyadic Deontic Logic by Carmo and Jones *)
4 typedecl 'i (*type for possible worlds*) type_synonym σ = "i⇒bool"
5 consts av:::"i⇒σ" pv:::"σ⇒(σ⇒bool)" (*accessibility relations*) cw:::i (*current world*)
6
7 axiomatization where
8   ax_3a: "∃x. av(w)(x)" and ax_4a: "∀x. av(w)(x) → pv(w)(x)" and ax_4b: "pv(w)(w)" and
9   ax_5a: "¬ob(X)(x). False)" and
10  ax_5b: "(∀w. ((Y(w) → X(w)) → (Z(w) ∧ X(w)))) → (ob(X)(Y) → ob(X)(Z))" and
11  ax_5c: "(∀Z. β(Z) → ob(X)(Z)) ∧ (∃Z. β(Z))" →
12    (((∃y. ((Aw. ∀Z. β(Z) → (Z(w) ∧ X(y))) → ob(X)(λw. ∀Z. (β Z) → (Z w))))" and
13    ax_5d: "((∀w. Y(w) → X(w)) ∧ ob(X)(Y) ∧ (∀w. X(w) → Z(w)))" →
14      → ob(Z)(λw. (Z(w) ∧ ¬X(w)) ∨ Y(w))" and
15  ax_5e: "((∀w. Y(w) → X(w)) ∧ ob(X)(Z) ∧ (∃w. Y(w) ∧ Z(w))) → ob(Y)(Z)" and
16
17 abbreviation ddlneg ("¬_[52]53) where "¬A ≡ λw. ¬A(w)"
18 abbreviation ddland (infixr "∧" 51) where "A ∧ B ≡ λw. A(w) ∧ B(w)"
19 abbreviation ddlor (infixr "∨" 50) where "A ∨ B ≡ λw. A(w) ∨ B(w)"
20 abbreviation ddlimp (infixr "→" 49) where "A → B ≡ λw. A(w) → B(w)"
21 abbreviation ddlequiv (infixr "↔" 48) where "A ↔ B ≡ λw. A(w) ↔ B(w)"
22 abbreviation ddbox ("□") where "□A ≡ λw. ∀v. A(v)" (*A = (λw. True)*)
23 abbreviation ddboxa ("□a") where "□a ≡ λw. (λv. av(w)(x) → A(x))" (*in all actual worlds*)
24 abbreviation ddboxp ("□o") where "□o ≡ λw. (λv. pv(w)(x) → A(x))" (*in all potential worlds*)
25 abbreviation ddldia ("◇") where "◇A ≡ ▦(¬A)"
26 abbreviation ddldlia ("◇a") where "◇a ≡ ▦(¬a)"
27 abbreviation ddldipa ("◇o") where "◇o ≡ ▦(¬o(¬A))"
28 abbreviation ddlo ("O[_)") [52]53) where "O[B|A] ≡ λw. ob(A)(B)" (*it ought to be ψ, given φ*)
29 abbreviation ddloa ("Oa") where "Oa ≡ λw. ob(av(w))(A) ∧ (∃x. av(w)(x) ∧ ¬A(x))" (*actual obligation*)
30 abbreviation ddlop ("Oo") where "Oo ≡ λw. ob(pv(w))(A) ∧ (∃x. pv(w)(x) ∧ ¬A(x))" (*primary obligation*)
31 abbreviation ddltop ("T") where "T ≡ λw. True"
32 abbreviation ddbot ("⊥") where "⊥ ≡ λw. False"
33
34 abbreviation ddvalid:::"σ ⇒ bool" ("_|" [7]105) where "|A| ≡ λw. A w" (*Global validity*)
35 abbreviation ddlidcw:::"σ ⇒ bool" ("_|cw" [7]105) where "|A|cw ≡ A cw" (*Local validity (in cw)*)
36
37 (* A is obligatory *)
38 abbreviation obligatoryDDL:::"σ ⇒ σ" ("O(_)") where "O(A) ≡ O(A|T)"
39
40 (* Consistency *)
41 lemma True nitpick [satisfy] oops

```

Documentation Sidebar State Theories

Output Query Sledgehammer Symbols

The screenshot shows the Isabelle/HOL IDE interface with the file `GDPR.thy` open. The code defines obligations related to data processing lawfully, erasing data, and killing a boss, and includes experiments with nitpick and sledgehammer.

```

1 theory GDPR imports SDL (* Christoph Benzmueller & Xavier Parent, 2018 *)
2
3 begin (** GDPR Example **)
4 consts process_data_lawfully::σ erase_data::σ kill_boss::σ
5
6 axiomatization where
7   (* It is an obligation to process data lawfully. *)
8   A1: "[0(process_data_lawfully)]" and
9   (* Implicit: It is an obligation to keep the data if it was processed lawfully. *)
10  Implicit: "[0(process_data_lawfully → ¬erase_data)]" and
11  (* If data was not processed lawfully, then it is an obligation to erase the data. *)
12  A2: "[¬process_data_lawfully → 0(erase_data)]"
13  (* Given a situation where data is processed unlawfully. *) and
14  A3: "[¬process_data_lawfully]_cw"
15
16 (** Some Experiments **)
17 lemma True nitpick [satisfy] oops (* Consistency-check: Is there a model? *)
18 lemma False sledgehammer oops (* Inconsistency-check: Can Falsum be derived? *)
19
20 lemma "[0(erase_data)]" sledgehammer nitpick oops (* Should the data be erased? *)
21 lemma "[0(¬erase_data)]" sledgehammer nitpick oops (* Should the data be kept? *)
22 lemma "[0(kill_boss)]" sledgehammer nitpick oops (* Should the boss be killed? *)
23 end

```

The interface includes tabs for Documentation, Sidekick, State, and Theories, and a bottom panel with Proof state, Auto update, Update, Search, and a status message about sledgehammering and proof found.

Bottom navigation bar: Output, Query, Sledgehammer, Symbols

Demo I: Global vs. Local Consequence Relation

The screenshot shows the Isabelle/HOL IDE interface. The top part displays the theory file `GDPRGlobal.thy` with its imports, global obligations, and experiments. The bottom part shows the proof process for a lemma involving the `sledgehammer` command.

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

1 theory GDPRGlobal imports DDL      (* Christoph Benzmueller & Xavier Parent, 2018 *)
2
3 begin (** GDPR Example **)
4 consts process_data_lawfully::σ erase_data::σ kill_boss::σ
5
6 axiomatization where
7   (* It is an obligation to process data lawfully. *)
8   A1: "[0(process_data_lawfully)]"
9   (* Given a situation where data is processed unlawfully. *) and
10  A3: "[~process_data_lawfully]"
11
12 (** Some Experiments **)
13 lemma True nitpick [satisfy] nunchaku [satisfy] oops (* Consistency-check: Is there a model? *)
14 lemma False sledgehammer oops (* Inconsistency-check: Can Falsum be derived? *)
15
16 lemma "[0(erase_data)]" sledgehammer nitpick oops (* Should the data be erased? *)
17 lemma "[0(~erase_data)]" sledgehammer nitpick oops (* Should the data be kept? *)
18 lemma "[0(kill_boss)]" sledgehammer nitpick oops (* Should the boss be killed? *)
19 end
20
21
22
23
```

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

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!

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

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

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

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

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

```
(*I0 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]

The screenshot shows the Isabelle/HOL IDE interface with the theory file `IO_Logic.thy` open. The code contains several lemmas and abbreviations related to I/O-Logic, specifically GDPR examples. The interface includes a toolbar with various icons, a navigation bar with tabs like "Output", "Query", "Sledgehammer", and "Symbols", and a sidebar with tabs for "Documentation", "Sidekick", "State", and "Theories".

```
(* Some Tests *)
consts a::e b::e e::e
abbreviation "G1 ≡ (λX. X=(a,e) ∨ X=(b,e))" (* G = {(a,e),(b,e)} *)
lemma "outl G1 a e" by blast (*proof*)
lemma "outpre G1 a e" by blast (*proof*)
lemma "outpre G1 (a ∨ b) e" nitpick oops (*countermodel*)
lemma "outl G1 (a ∨ b) e" nitpick oops (*countermodel*)
lemma "[x] ==> outpre G1 (a ∨ b) x" nitpick oops (*countermodel*)
lemma "[x] ==> outl 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 "outl G (¬pr_d_lawf) erase_d" by smt (*proof*)
lemma "outl G (¬pr_d_lawf) (¬erase_d)" nitpick oops (*countermodel*)
lemma "outl G (¬pr_d_lawf) kill_boss" nitpick oops (*countermodel*)
lemma "outl G (¬pr_d_lawf) ⊥" nitpick oops (*countermodel*)
```

Proof state Auto update Update Search: 100%

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

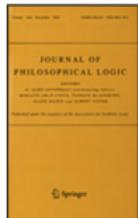
Skolem constant:
w = i₁

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

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...

The screenshot shows the Isabelle/HOL IDE interface. On the left, there's a vertical bar with the "JOURNAL OF PHILOSOPHICAL LOGIC" logo and a Springer logo. The main window displays the theory file `PrefDDL.thy` located at `~/chris/trunk/tex/talks/2018-DEON/DEMO/`. The code in the file includes various lemmas and definitions related to proof theory, soundness, and inference rules, such as `classical`, `OM`, `DFP`, `COK`, `SOK`, `abs`, `nec`, `ext`, `id`, `Sh`, and `MP`. It also contains a lemma `CM` which fails to prove due to nitpick. At the bottom, the nitpick results are shown:

```
Nitpicking formula...
Nitpick found a counterexample for card 'w' = 3:

Free variables:
  A = (λx. _)(w1 := True, w2 := True, w3 := True)
  B = (λx. _)(w1 := True, w2 := True, w3 := False)
  C = (λx. _)(w1 := False, w2 := True, w3 := False)
  op ≥ = (λx. _)
    (w1 := (λx. _)(w1 := True, w2 := True, w3 := False),
     w2 := (λx. _)(w1 := True, w2 := True, w3 := True),
     w3 := (λx. _)(w1 := True, w2 := False, w3 := True))
  opt = (λx. _)
```

The interface includes tabs for Output, Query, Sledgehammer, and Symbols, and a status bar showing Proof state, Auto update, Update, Search, and zoom level (100%).

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

The screenshot shows the Isabelle/HOL proof assistant interface. The top window displays the source code for `Gewirth3.thy`, which contains a proof of the principle of generic consistency. The proof involves several stages (I, II, IIIa, IIIb) and uses various rules like `impI`, `allI`, and `simp`. The bottom window shows the proof state with a single subgoal involving quantifiers and function application.

```
15 (*C9: "Every PPA has a necessary right to their freedom and well-being"*)
16 theorem C9: "[ $\forall a. \text{PPA } a \rightarrow \square.\text{RightTo } a \text{ FWB}$ ]"
17 proof -
18 {
19   fix I {
20     fix E {
21       (** Stage I *)
22       assume P1: "[ $\text{ActsOnPurpose } I E$ ]" (*I act voluntarily on purpose E*)
23       from P1 have P1_var: "[ $\text{PPA } I$ ]" by auto (*definition of PPA*)
24       from P1 have C2: "[ $\text{Good } I E$ ]" using explicationGoodness1 by blast (*E is good for me (I)*)
25       hence C4: "[ $\square.\text{Good } I (\text{FWB } I)$ ]" using explicationGoodness2 P3 by blast (*My F&WB are necessary goods*)
26       (** Stage II *)
27       hence C4a: "[ $\square(\text{FWB } I \mid \square.\text{Good } I (\text{FWB } I))$ ]" using explicationGoodness3 explicationFWB1 by blast
28       hence C4b: "[ $\square_1(\text{FWB } I)$ ]" using explicationFWB2 C4 CJ_14p by blast
29       hence C4c: "[ $\square_1(\square_2(\text{FWB } I))$ ]" using OIOAC by auto
30       hence C5a: "[ $\square_1(\neg\text{InterferesWith } a (\text{FWB } I))$ ]" using explicationInterference2 by auto
31       hence C5: "[ $\square.\text{RightTo } I \text{ FWB}$ ]" by simp (*I have a claim right to my freedom and well-being*)
32       hence C5_var: "[ $\square.\text{RightTo } I \text{ FWB}$ ]" by simp
33     }
34     (** Stage IIIa *)
35     hence C6: "[ $\text{ActsOnPurpose } I E \rightarrow \square.\text{RightTo } I \text{ FWB}$ ]" by (rule impI)
36   }
37   hence C7: "[ $\forall P. \text{ActsOnPurpose } I P \rightarrow \square.\text{RightTo } I \text{ FWB}$ ]" by (rule allI)
38 }
39 hence CB: "[ $\forall a. \forall P. \text{ActsOnPurpose } a P \rightarrow \square.\text{RightTo } a \text{ FWB}$ ]" by (rule allI)
40 hence C9_var: "[ $\forall a. \text{PPA } a \rightarrow \square.\text{RightTo } a \text{ FWB}$ ]"
41 by simp (*Every PPA has a necessary right to their freedom and well-being*)
42 thus ?thesis by simp
43 qed
```

Proof state:

```
proof (prove)
goal (1 subgoal):
  1. ( $\lambda x. [\text{PPA } x]$ )  $\sqsubseteq (\lambda x. \text{pv aw} \sqsubseteq \square_1(\lambda w. \forall x_a. (\neg\text{InterferesWith } x_a (\text{FWB } x)) w))$ 
```

Output Query Sledgehammer Symbols

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>



$\forall P$.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>

 $\forall P$.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>

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	\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)



ASCII	Syntax	Meaning
~	\neg	Negation
	\vee	Disjunction
&	\wedge	Conjunction
=>	\Rightarrow	Material implication
\Leftrightarrow	\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>).`

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, $0(processDataLawfully)).  
ddl(a2, axiom, $0(eraseData/~processDataLawfully)).  
ddl(a3, localAxiom, ~processDataLawfully).  
  
ddl(c1, conjecture, $0(eraseData)).
```

... giving ...

```
% SWS 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, $0(processDataLawfully)).  
ddl(a2, axiom, $0(eraseData/~processDataLawfully)).  
ddl(a3, localAxiom, ~processDataLawfully).  
  
ddl(c1, conjecture, $0(eraseData)).
```

... giving ...

```
% SWS 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]

```
leopard:Leo3 cbenzmueller$ more gdpr.p
ddl(a1, axiom, $0(processDataLawfully)).
ddl(a2, axiom, (~processDataLawfully) $\rightarrow$  $0(eraseData)).
ddl(a3, localAxiom, ~processDataLawfully).

ddl(c1, conjecture, $0(eraseData)).

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

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.

Ethics

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

Ethics

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:



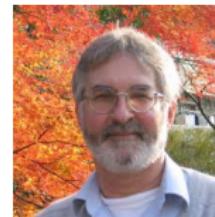
B. Woltzenl.-P.
(ANU Canberra)



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