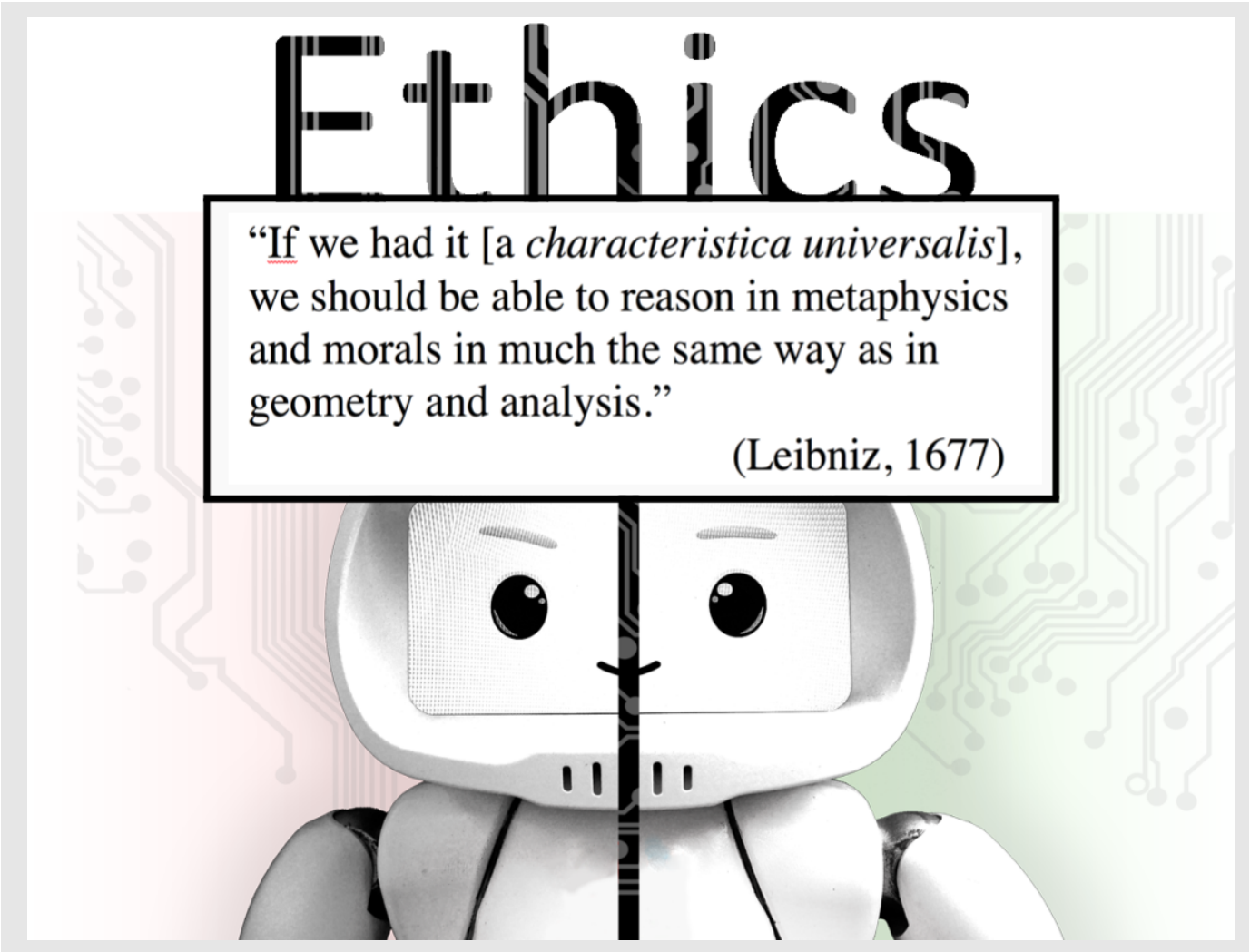
Universal Logical Reasoning
(with Applications in Normative Reasoning)

Christoph Benzmüller
Freie Universität Berlin & Latentine GmbH

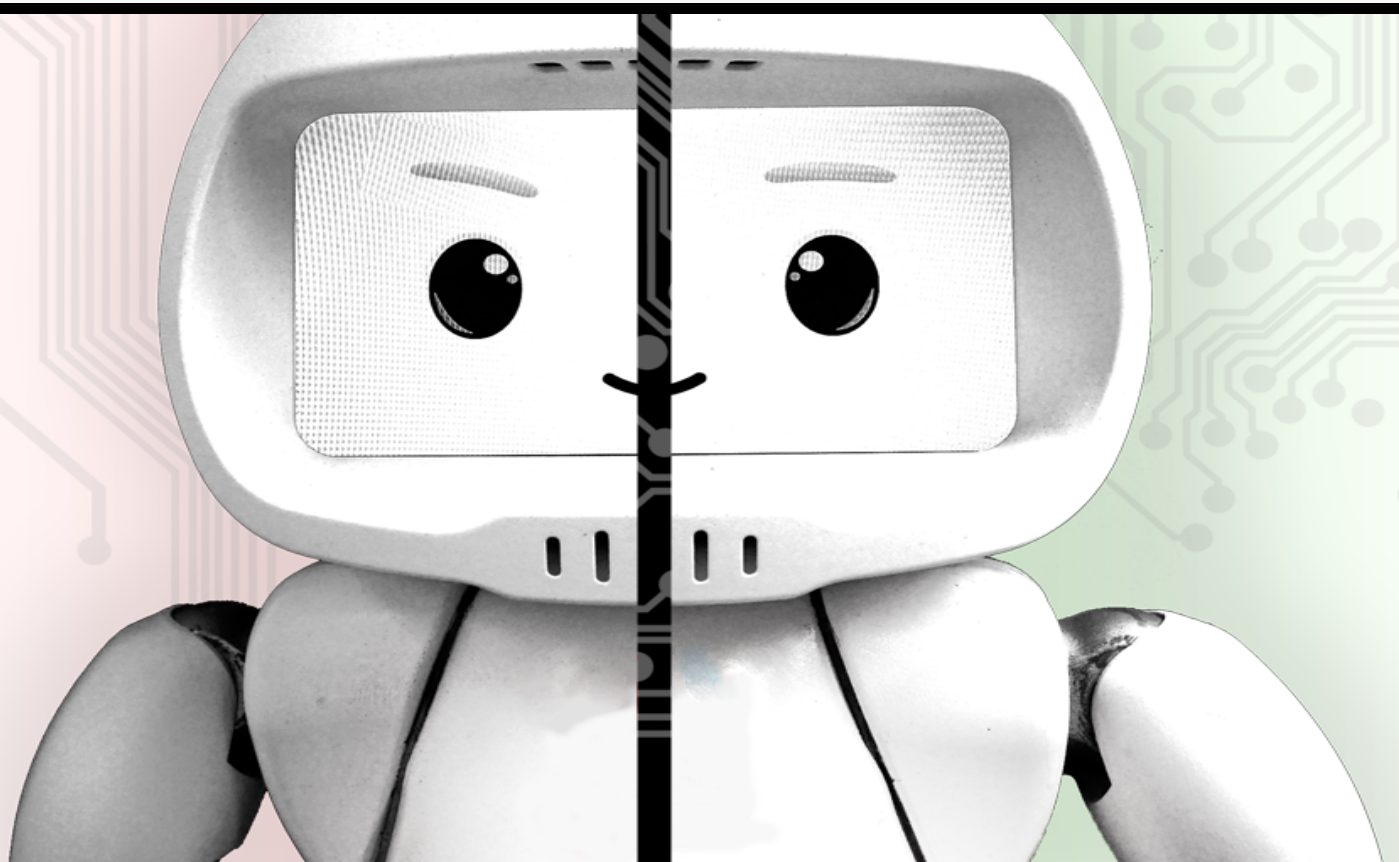


Logics in AI, Zheijang University — Oct 22, 2020

Ethics

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

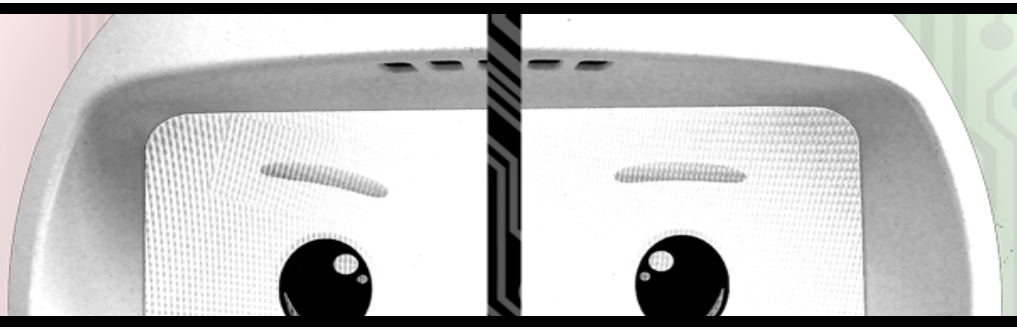
- appropriate forms of **machine-control**
- appropriate forms of **human-machine-interaction**



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?

Motivation: (Pseudo-)Ethical Intelligent Systems

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

- How to prevent Superintelligence from turning against humanity?

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

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

Different kinds of systems and approaches:

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

Motivation: (Pseudo-)Ethical Intelligent Systems

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

- How to prevent Superintelligence from turning against humanity?

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

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

Different kinds of systems and approaches:

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

Motivation: (Pseudo-)Ethical Intelligent Systems

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

- How to prevent Superintelligence from turning against humanity?

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

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

Different kinds of systems and approaches:

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

Motivation: (Pseudo-)Ethical Intelligent Systems

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

- How to prevent Superintelligence from turning against humanity?

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

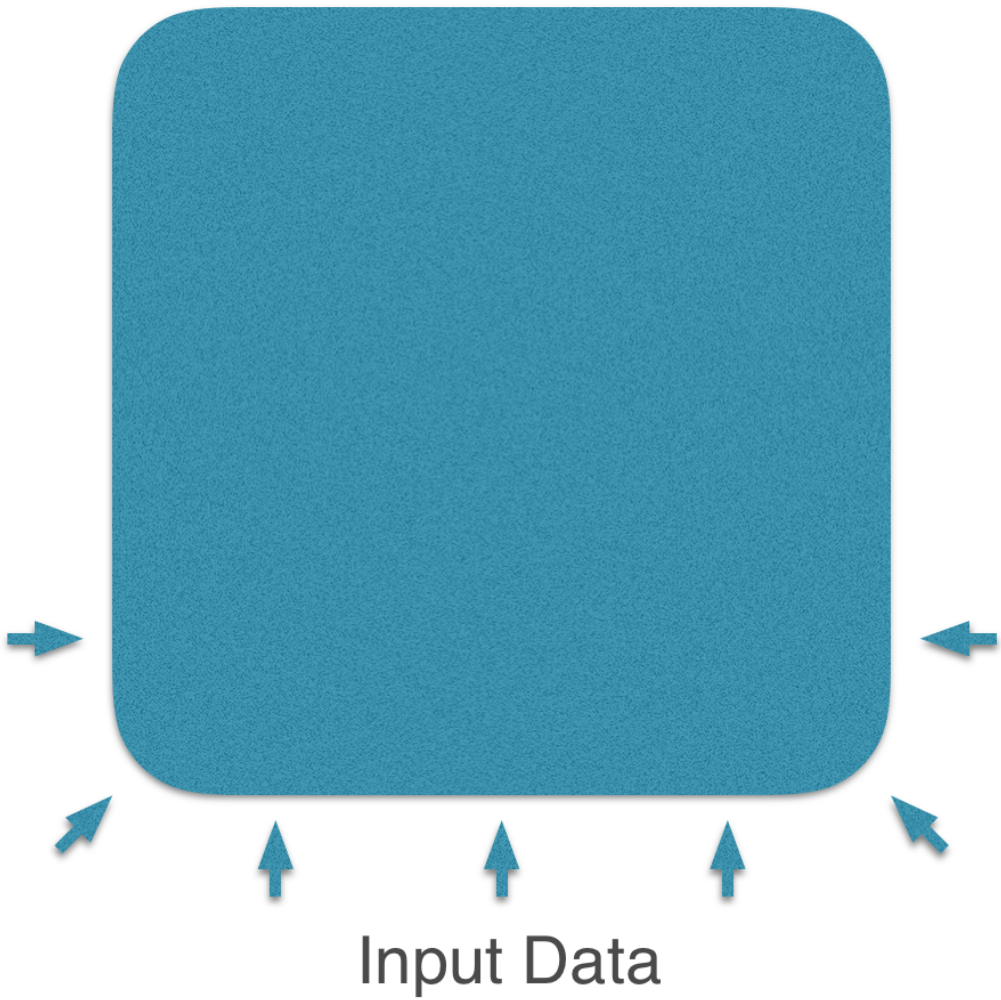
- Which norms? Which reasoning principles?
- What architectural design? What functionalities?
- How to implement, deploy and verify?

Different kinds of systems and approaches:

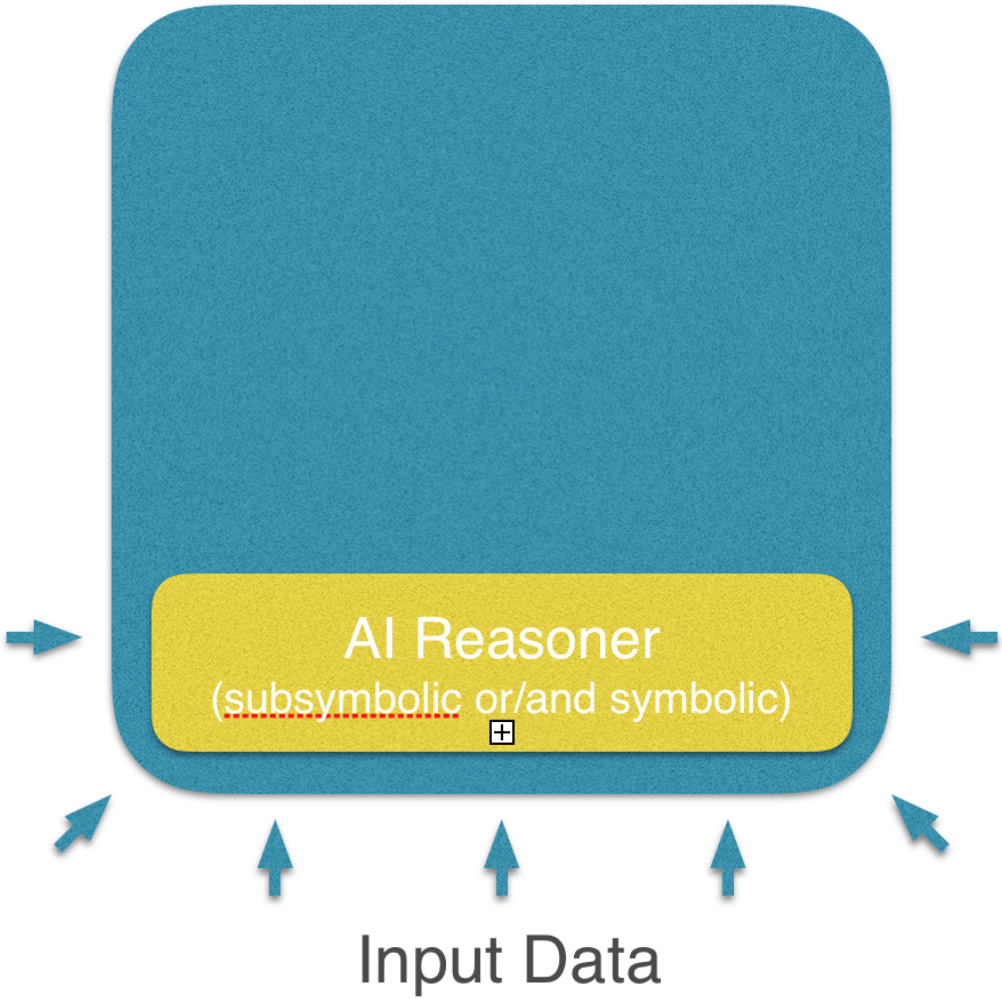
- [Moor, 2009]:
  - ethical impact agents (ethical consequences to actions)
  - implicit ethical agents (ethical reactions to given situations)
  - explicit ethical agents (reasoning with ethical theories/rules)
  - full ethical agents (conscious, intentional, free will)
- bottom-up vs. top-down
- [DoranEtAl.,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

(Pseudo-)Ethical Intelligent Systems

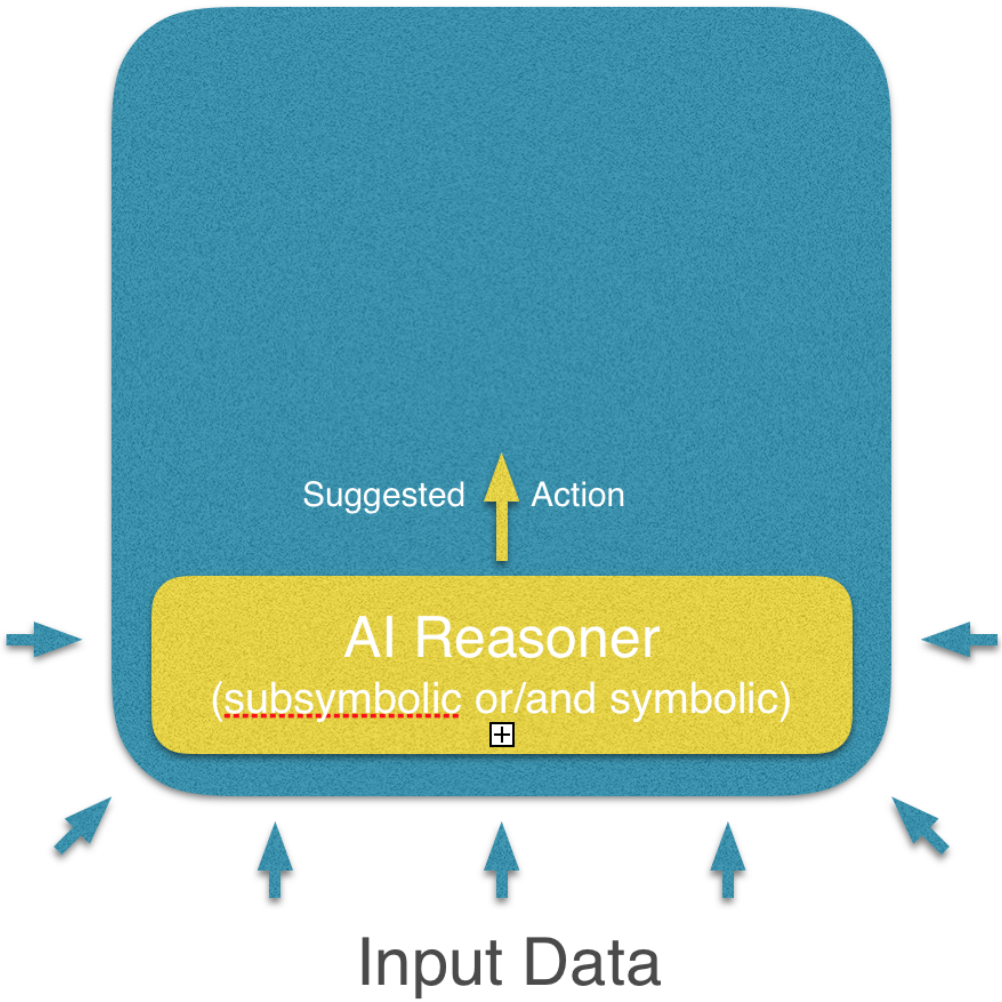
(Pseudo-)Ethical Intelligent Systems



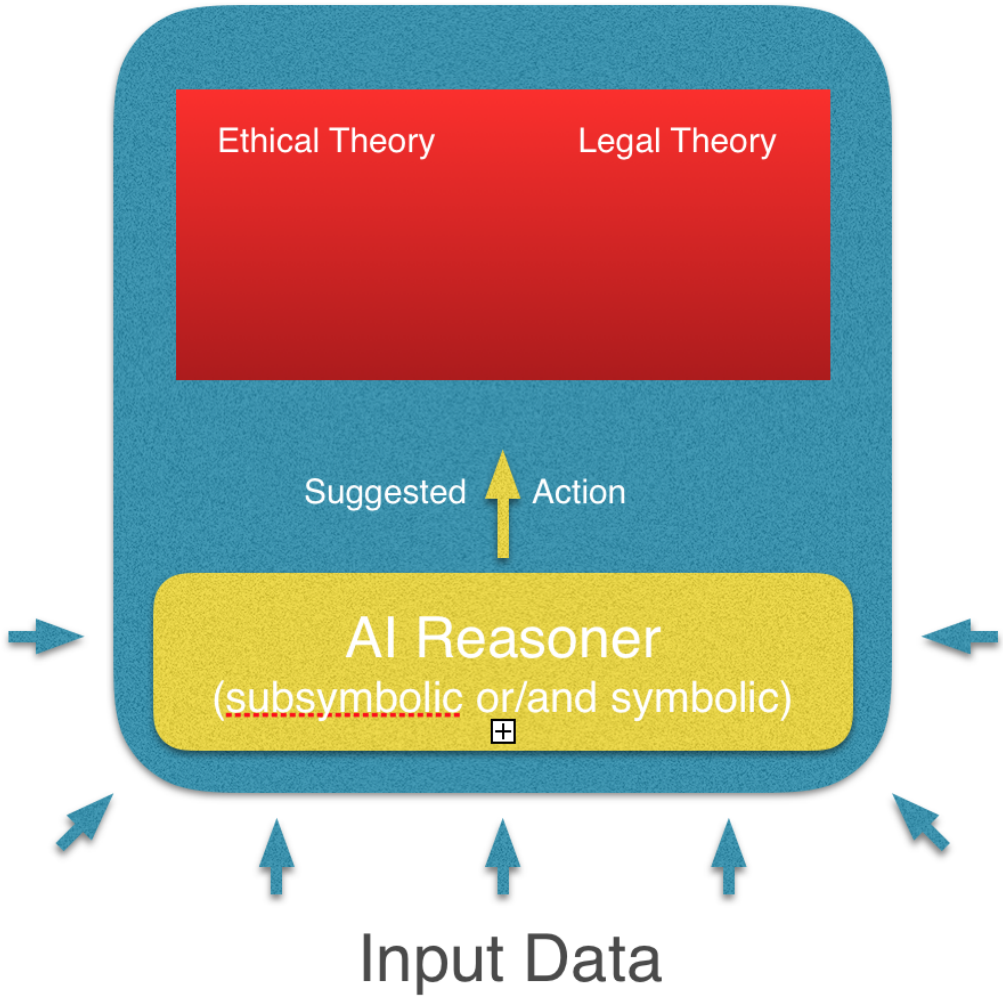
(Pseudo-)Ethical Intelligent Systems



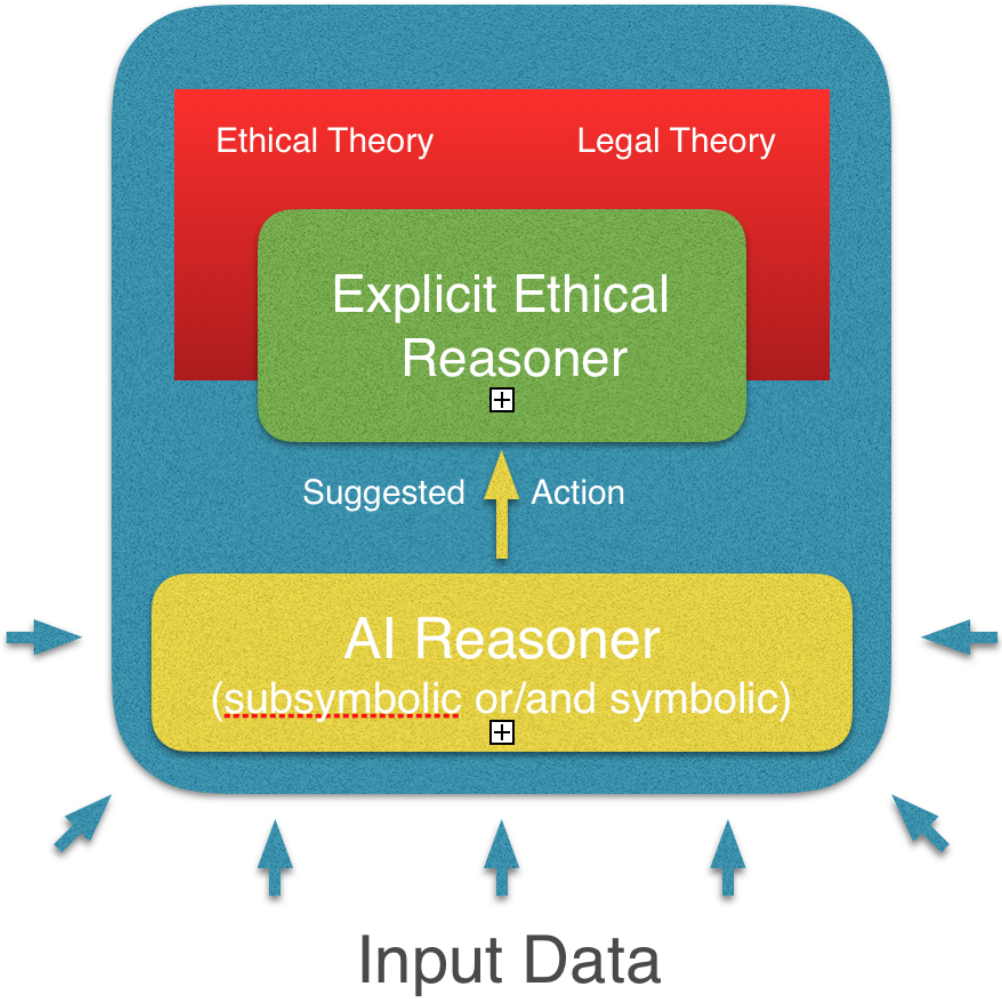
(Pseudo-)Ethical Intelligent Systems



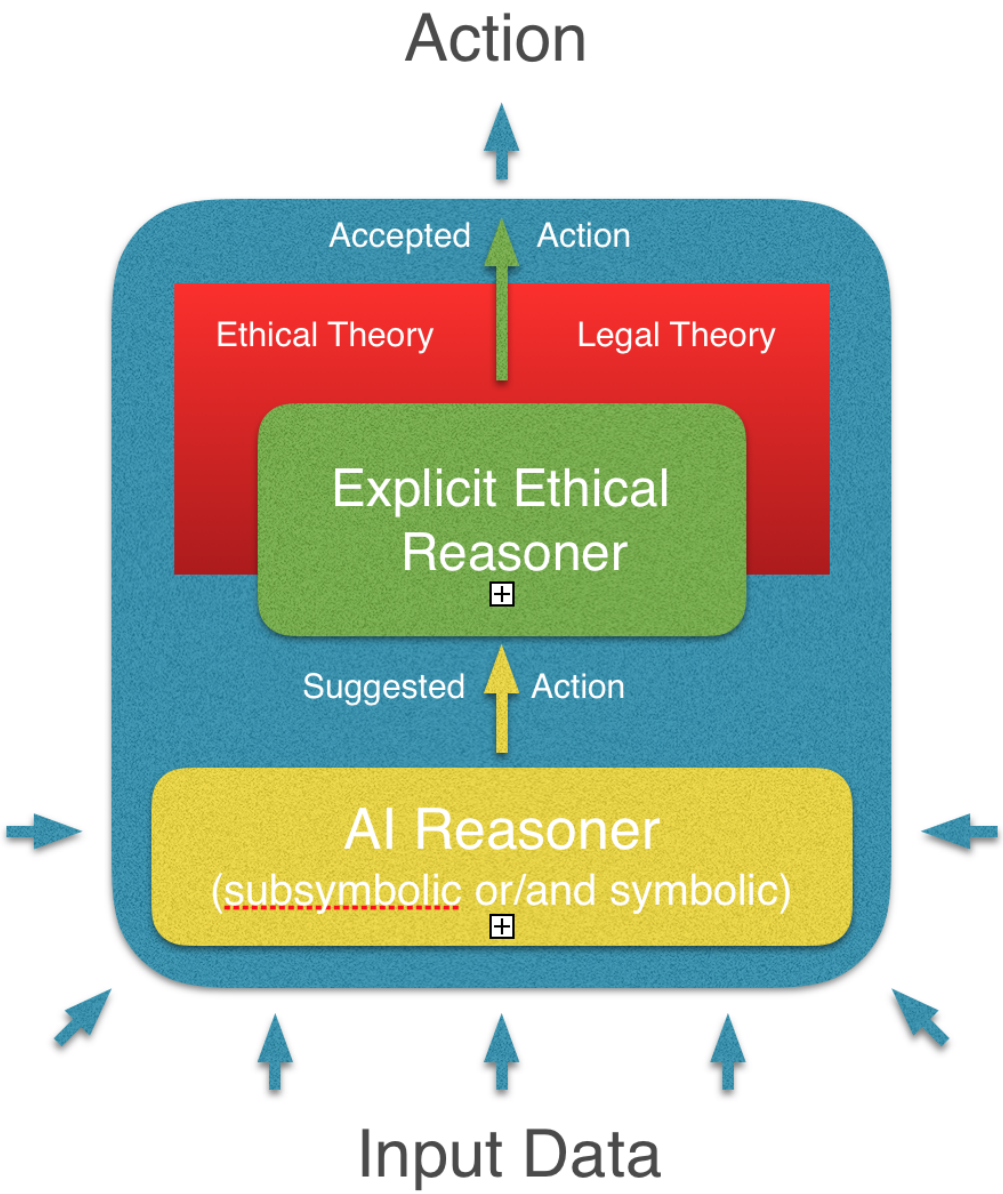
(Pseudo-)Ethical Intelligent Systems



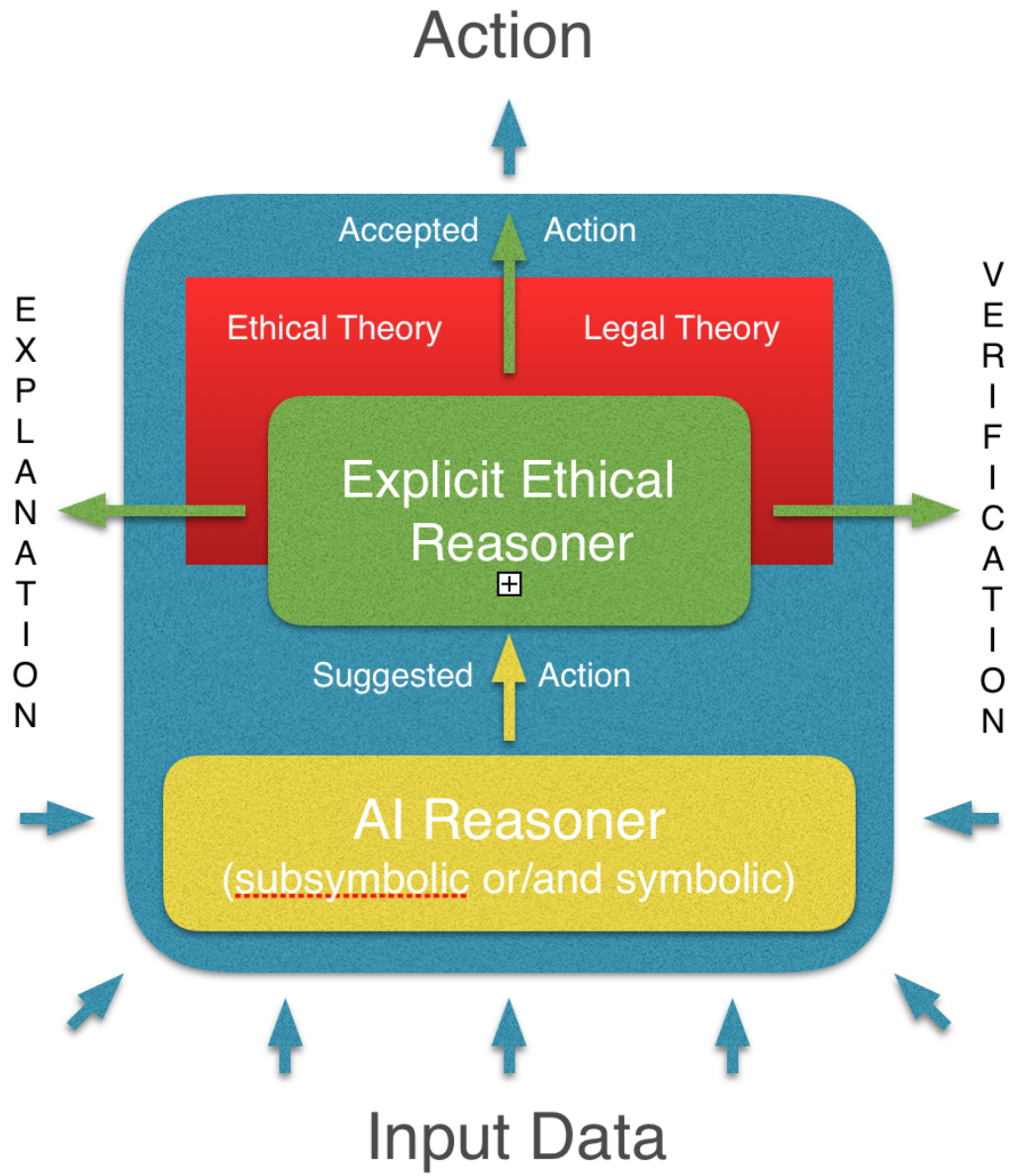
(Pseudo-)Ethical Intelligent Systems



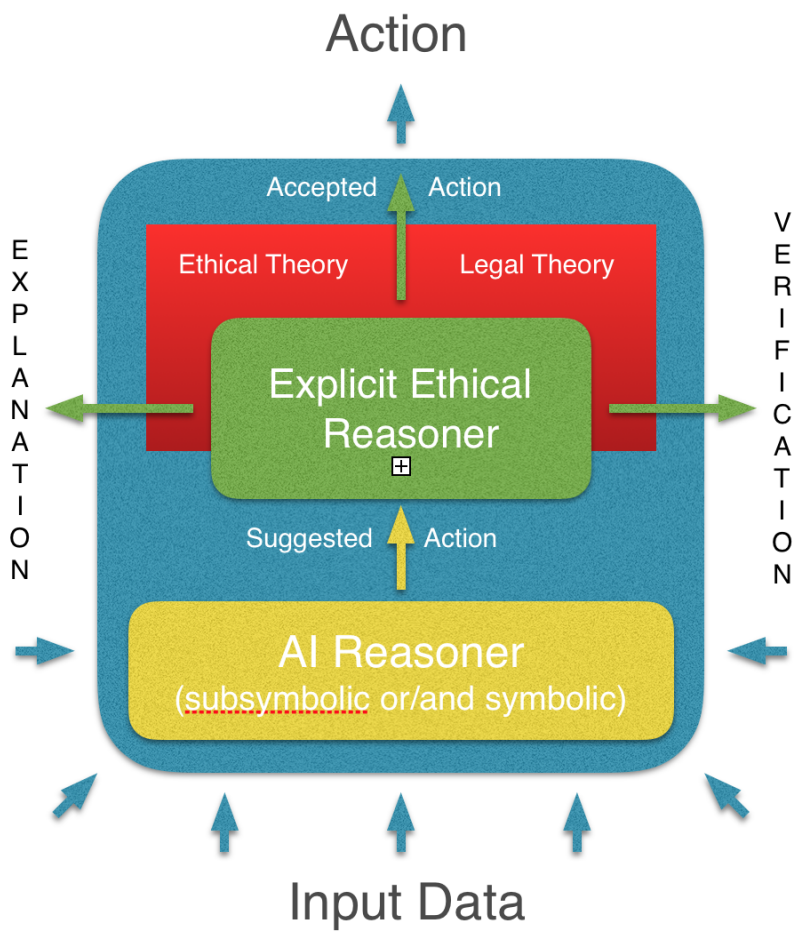
(Pseudo-)Ethical Intelligent Systems



(Pseudo-)Ethical Intelligent Systems



(Pseudo-)Ethical Intelligent Systems



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]
- . . .

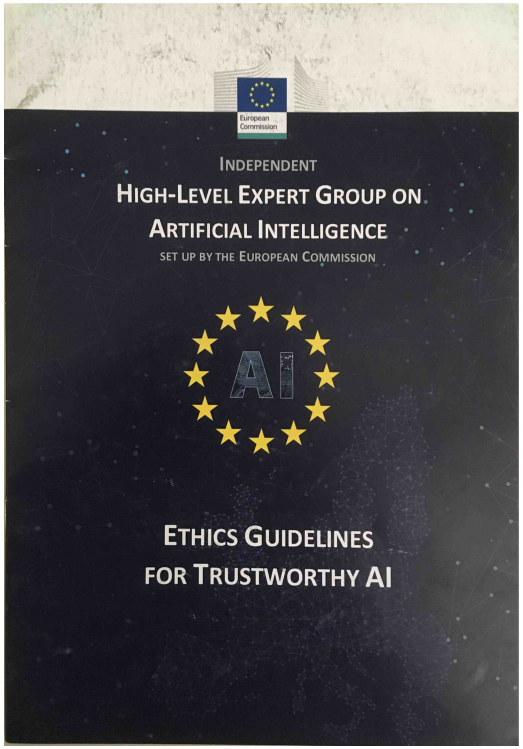
(Pseudo-)Ethical Intelligent Systems

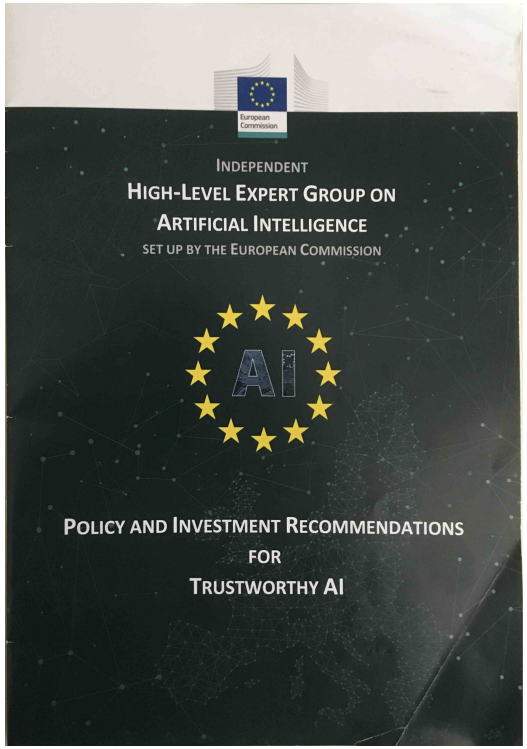
Address requests towards **Thrustworthy & Responsible AI**

- **German National AI Strategy (Nov 2018)**

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

- **Recommendations from the EU HLEG group (May & June 2019)**



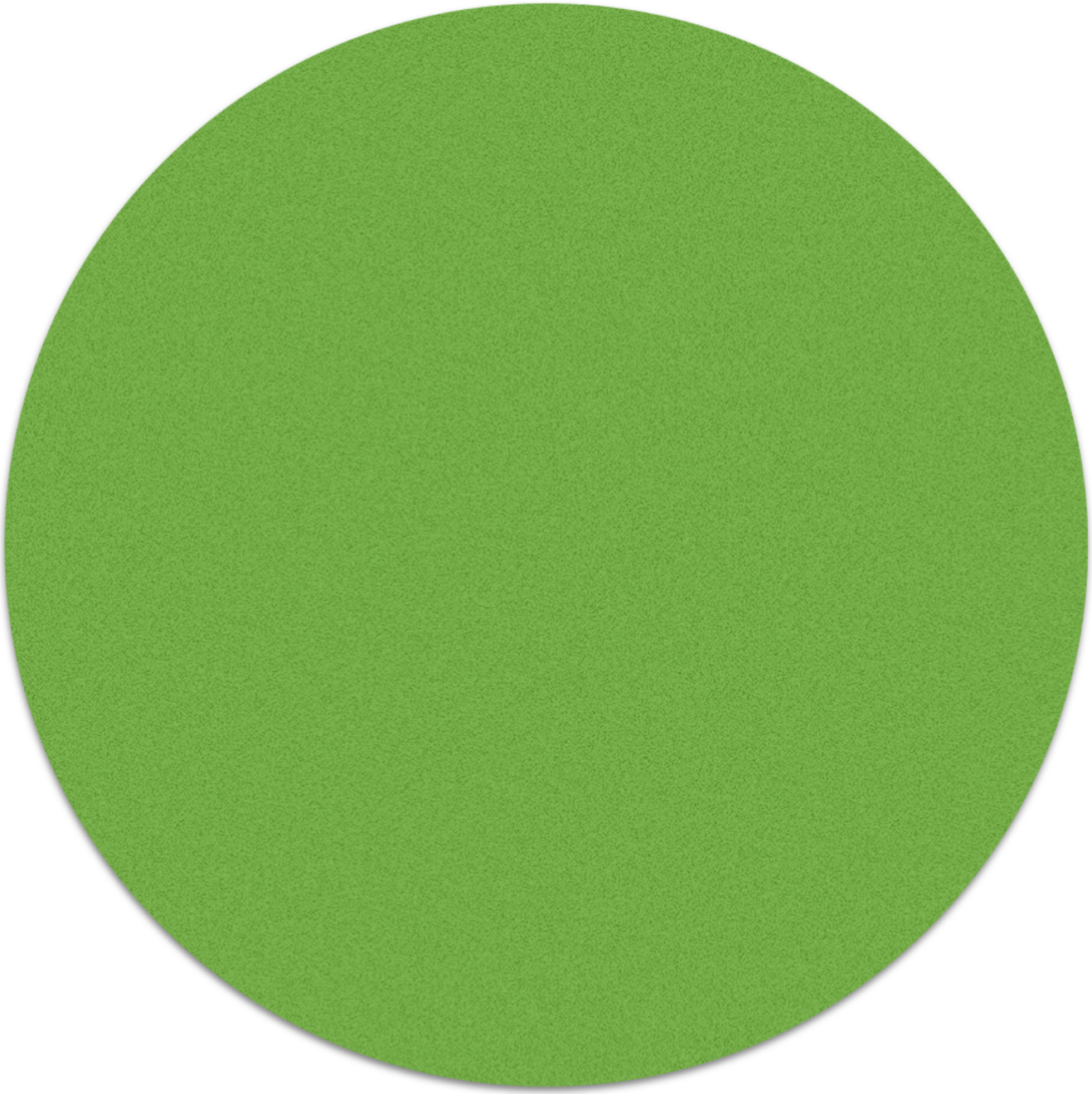


`https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence`

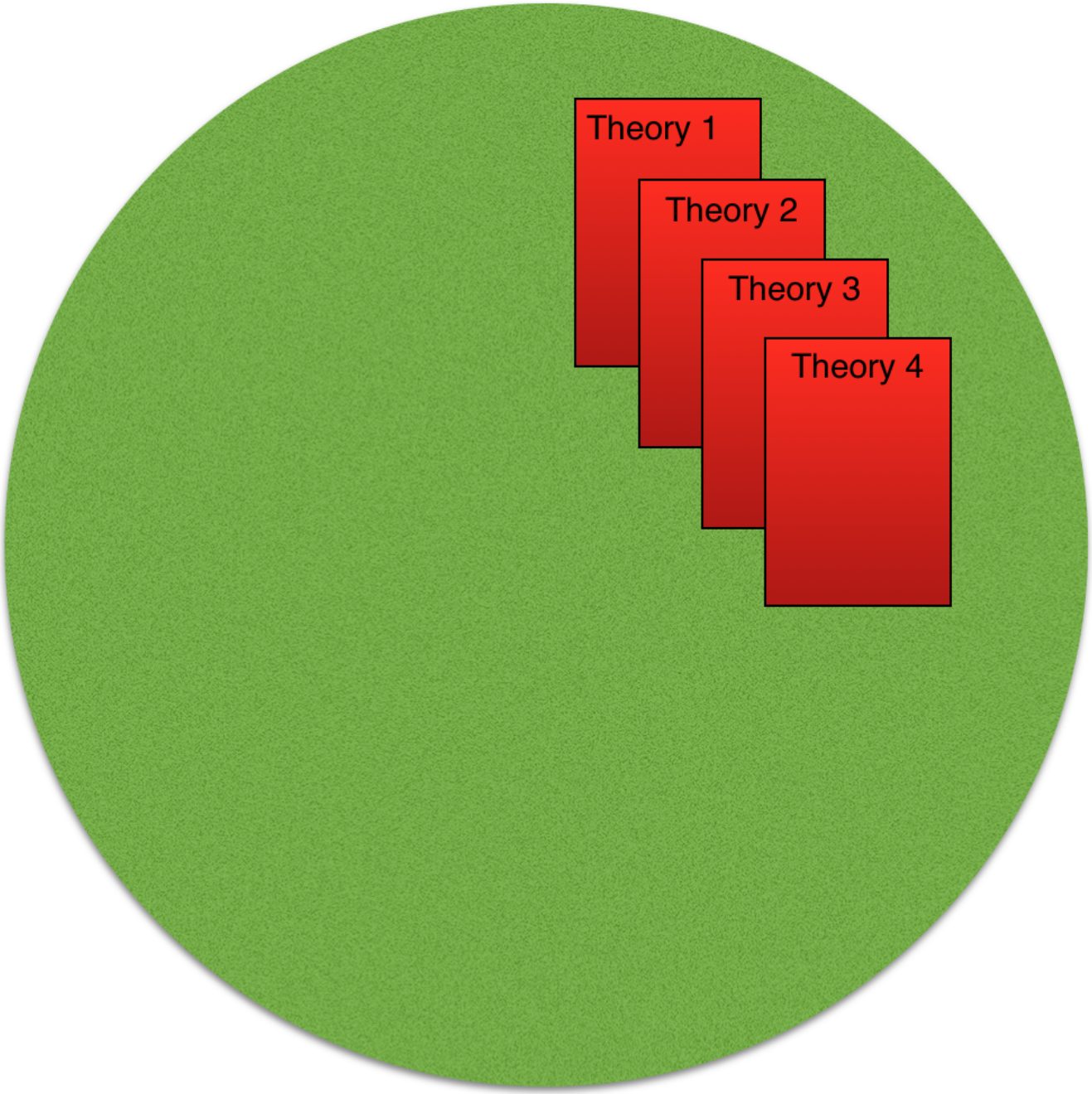
- **Ben Goertzel (CEO SingularityNET; Nov 2018): “Toward Democratic, Lawful Citizenship for AIs, Robots, and Corporations”**

  `https://tinyurl.com/y8h94ouv`

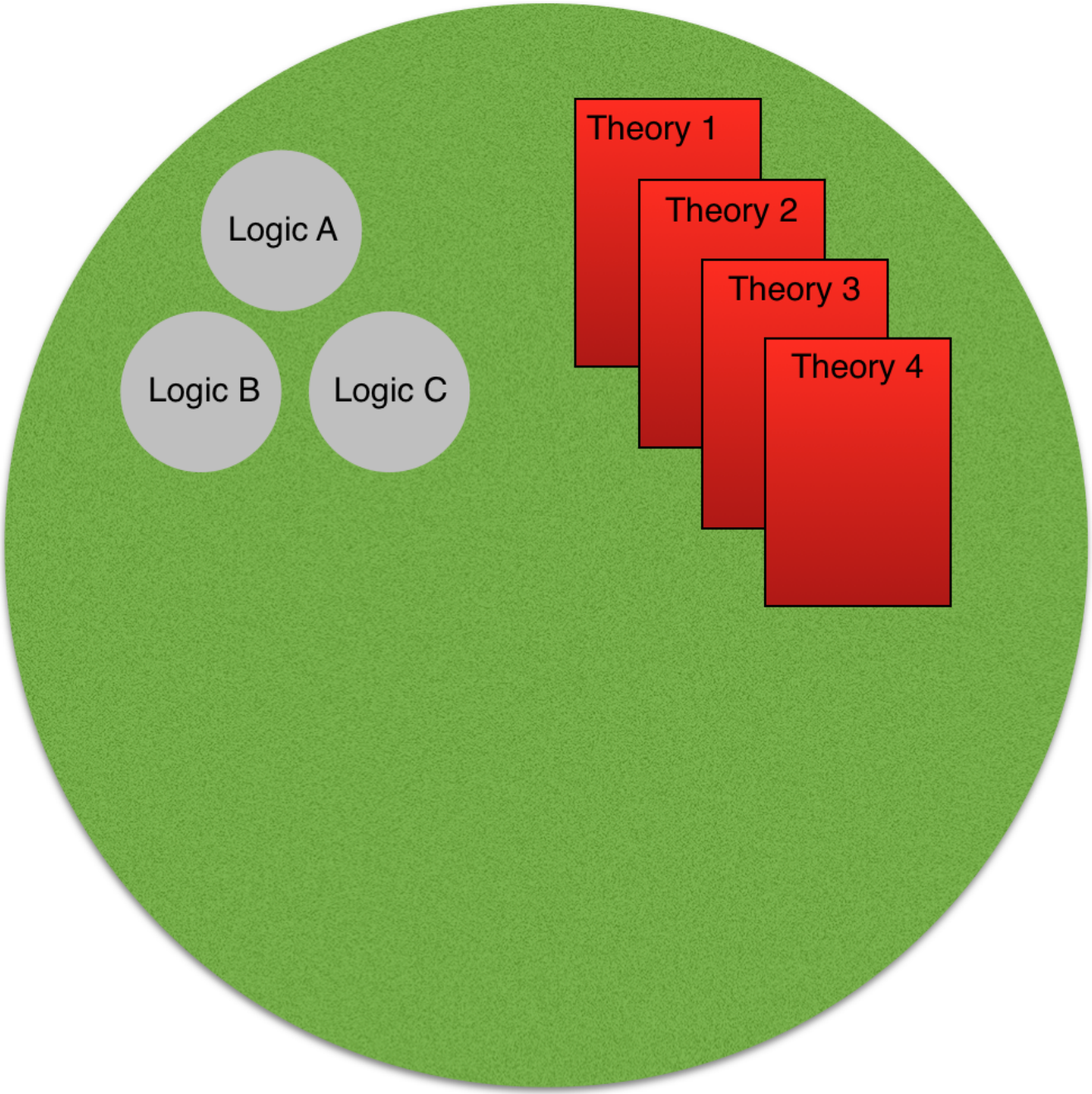
Normative Reasoning Experimentation Platform



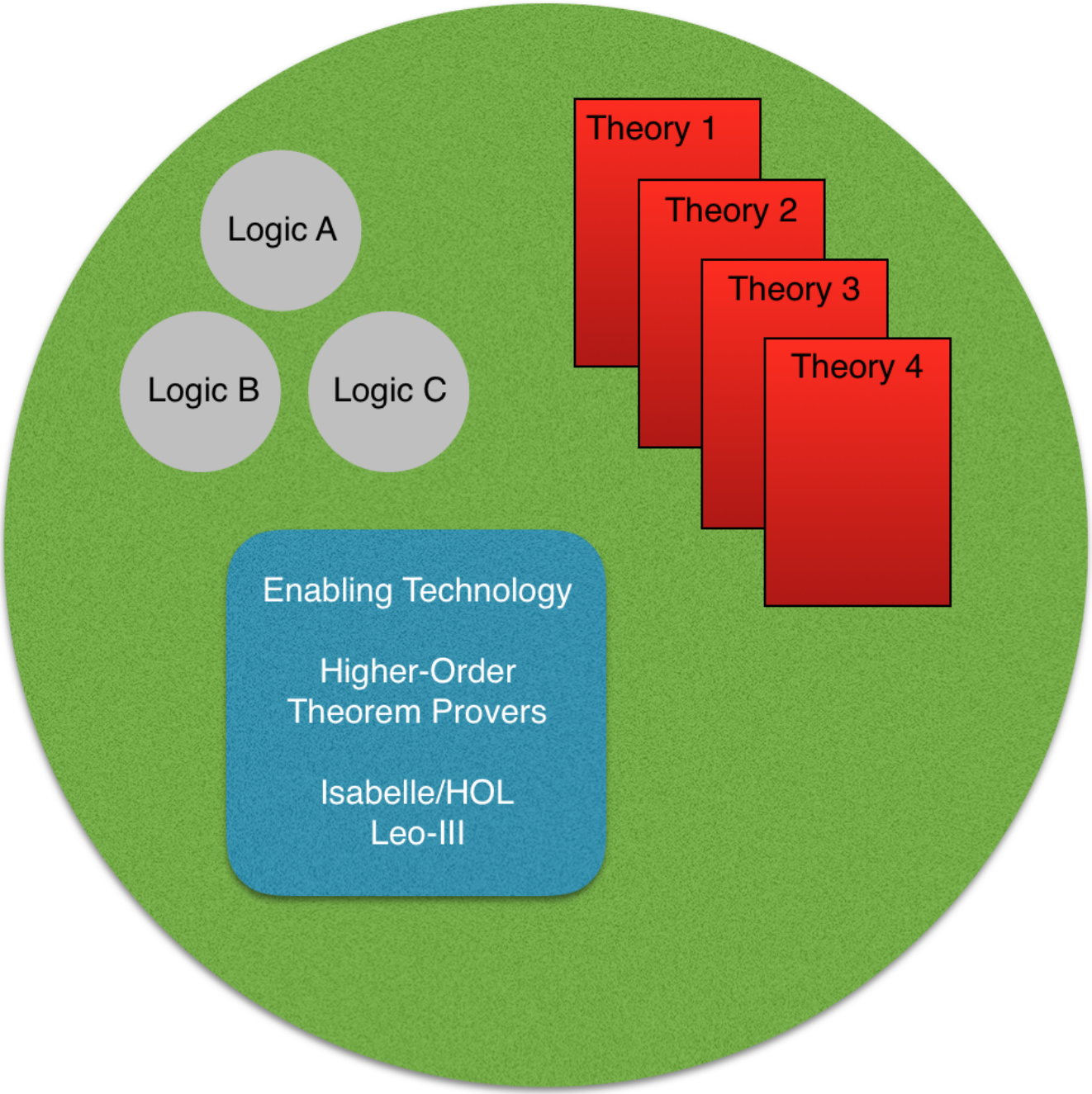
Normative Reasoning Experimentation Platform



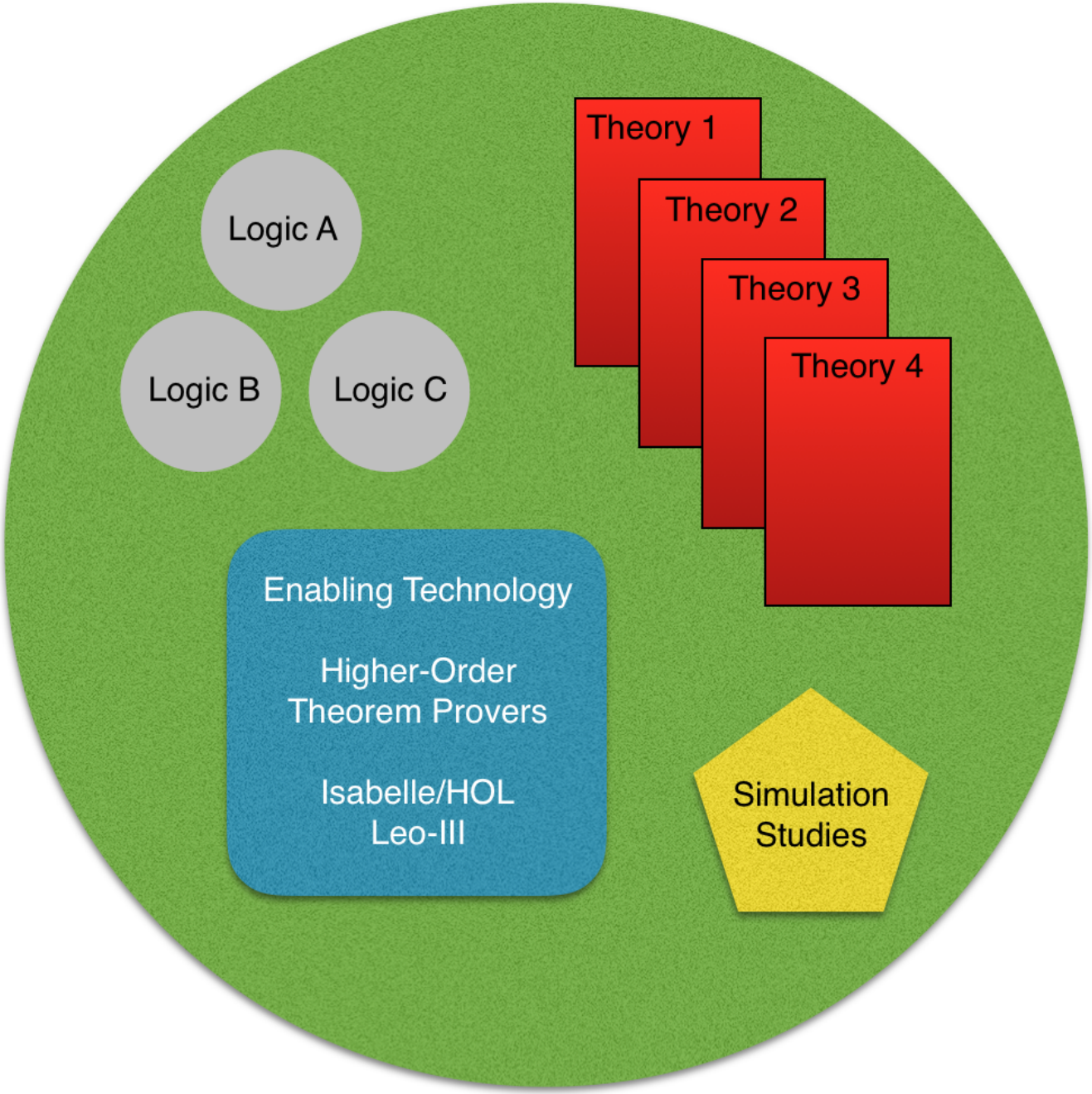
Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform



"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)

Flexible Reasoning Technology:
Shallow Semantical Embeddings (SSEs)
in Classical Higher-Order Logic (HOL)

Flexible Reasoning Technology: SSEs in HOL

Home | Video | Themen | Forum | English | DER SPIEGEL | SPIEGEL TV | Abo | Shop RSS | Mobile | Newsletter

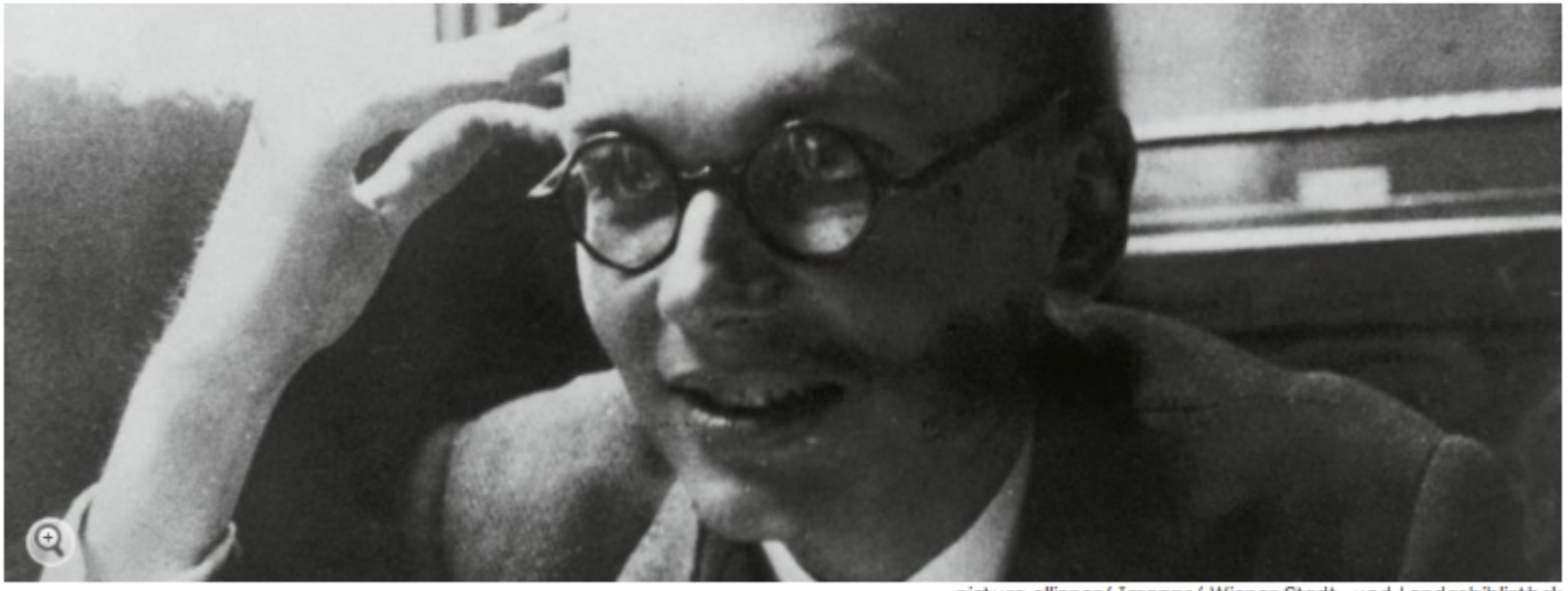
SPIEGEL ONLINE INTERNATIONAL Sign in | Register

Front Page World Europe Germany Business Zeitgeist Newsletter

English Site > Germany > Science > Scientists Use Computer to Mathematically Prove Gödel God Theorem

Holy Logic: Computer Scientists 'Prove' God Exists

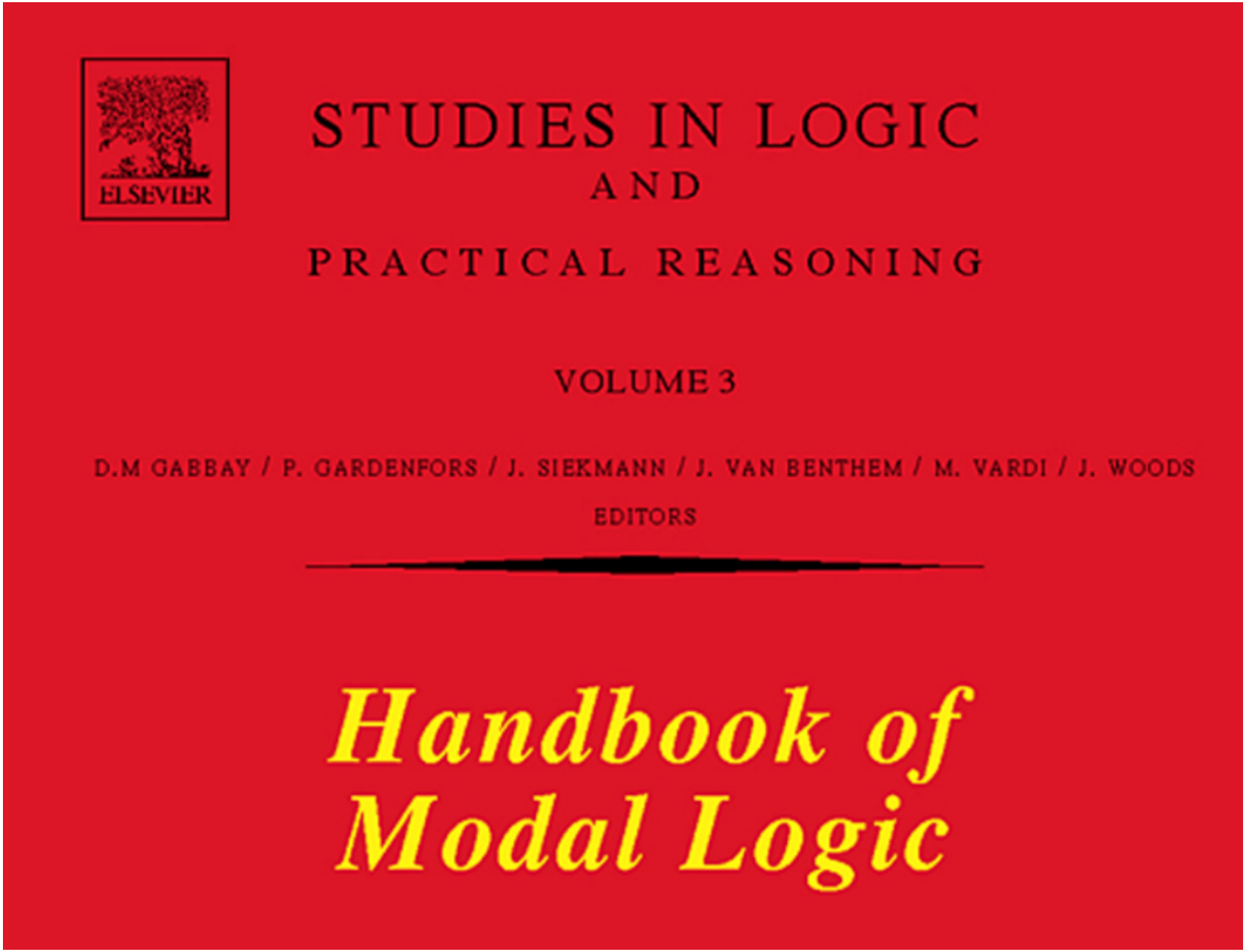
By David Knight



picture-alliance/ Imagno/ Wiener Stadt- und Landesbibliothek

Austrian mathematician Kurt Gödel kept his proof of God's existence a secret for decades. Now two scientists say they have proven it mathematically using a computer.

Two scientists have formalized a theorem regarding the existence of God penned by mathematician Kurt Gödel. But the God angle is somewhat of a red herring -- the real step forward is the example it sets of how computers can make scientific progress simpler.


ELSEVIER
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:

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

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

2 BASIC MODAL LOGIC

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

2.1 First steps in relational semantics

Syntax

Metalanguage

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

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

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

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,
$\mathfrak{M}, w \models \bot$		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 $R^m wv$ we have $\mathfrak{M}, v \models \varphi$,
$\mathfrak{M}, w \models [m] \varphi$	iff	for all $v \in W$ such that $R^m wv$ we have $\mathfrak{M}, v \models \varphi$.

A *model* (or *Kripke model*) $\mathfrak{M}$ for the basic modal language (over some fixed signature) is a triple $\mathfrak{M} = (W, \{R^m\}_{m \in \text{MOD}}, V)$. Here W, the *domain*, is a non-empty set, whose elements we usually call *points*, but which, for reasons which will soon be clear, are sometimes called *states*, *times*

in a model is a binary relation on W,

and

position symbol p in PROP a subset

$V(p)$

p is true. The first two components

$(W, \{$

model. If there is only one relation

in the

(W, R, V) for the model itself. We

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

Metalanguage

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

Semantics

$\mathfrak{M}, w \models p$	iff	$w \in V(p)$,
$\mathfrak{M}, w \models \top$		always,
$\mathfrak{M}, w \models \bot$		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 $R^m wv$ we have $\mathfrak{M}, v \models \varphi$,
$\mathfrak{M}, w \models [m] \varphi$	iff	for all $v \in W$ such that $R^m wv$ we have $\mathfrak{M}, v \models \varphi$.

Kripke Style Semantics

$M, g, s \models P$ if and only if $s \in g(P)$

$M, g, s \models \neg\, \varphi$ if and only if $M, g, s \not\models \varphi$

$M, g, s \models \varphi \vee \psi$ if and only if $M, g, s \models \varphi$ or $M, g, s \models \psi$

$M, g, s \models \Box\, \varphi$ if and only if for all t with sRt we have $M, g, t \models \varphi$

Kripke Style Semantics

$M, g, s \models P$ if and only if $s \in g(P)$

$M, g, s \models \neg\, \varphi$ if and only if $M, g, s \not\models \varphi$

$M, g, s \models \varphi \vee \psi$ if and only if $M, g, s \models \varphi$ or $M, g, s \models \psi$

$M, g, s \models \Box\, \varphi$ if and only if for all t with sRt we have $M, g, t \models \varphi$

Standard Translation for Propositional Fragment (encoded in HOL)

- "lifted" predicate $P_{i\to o}$
- $\neg_{(i\to o)\to(i\to o)}\ \varphi_{i\to o}\ w_i = \neg(\varphi w)$
- $\vee_{(i\to o)\to(i\to o)\to(i\to o)}\ \varphi_{i\to o}\ \psi_{i\to o}\ w_i = \varphi w \vee \psi w$
- $\Box_{(i\to o)\to(i\to o)}\ \varphi_{i\to o}\ w_i = \forall v_i\ Rwv \to \varphi v$

Kripke Style Semantics

$M, g, s \models P$ if and only if $s \in g(P)$

$M, g, s \models \neg\, \varphi$ if and only if $M, g, s \not\models \varphi$

$M, g, s \models \varphi \vee \psi$ if and only if $M, g, s \models \varphi$ or $M, g, s \models \psi$

$M, g, s \models \Box\, \varphi$ if and only if for all t with sRt we have $M, g, t \models \varphi$

Standard Translation for Propositional Fragment (encoded in HOL)

- "lifted" predicate $P_{i\to o}$
- $\neg_{(i\to o)\to(i\to o)}\ \varphi_{i\to o} = \lambda w_i\ \neg(\varphi w)$
- $\vee_{(i\to o)\to(i\to o)\to(i\to o)}\ \varphi_{i\to o}\ \psi_{i\to o} = \lambda w_i\ \varphi w \vee \psi w$
- $\Box_{(i\to o)\to(i\to o)}\ \varphi_{i\to o} = \lambda w_i\ \forall v_i\ Rwv \to \varphi v$

Kripke Style Semantics

$M, g, s \models P$ if and only if $s \in g(P)$

$M, g, s \models \neg\, \varphi$ if and only if $M, g, s \not\models \varphi$

$M, g, s \models \varphi \vee \psi$ if and only if $M, g, s \models \varphi$ or $M, g, s \models \psi$

$M, g, s \models \Box\, \varphi$ if and only if for all t with sRt we have $M, g, t \models \varphi$

Standard Translation for Propositional Fragment (encoded in HOL)

- "lifted" predicate $P_{i\to o}$
- $\neg_{(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda w_i\ \neg(\varphi w)$
- $\vee_{(i\to o)\to(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda\psi_{i\to o}\ \lambda w_i\ \varphi w \vee \psi w$
- $\Box_{(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda w_i\ \forall v_i\ Rwv \to \varphi v$

Kripke Style Semantics

$M, g, s \models P$ if and only if $s \in g(P)$

$M, g, s \models \neg\, \varphi$ if and only if $M, g, s \not\models \varphi$

$M, g, s \models \varphi \vee \psi$ if and only if $M, g, s \models \varphi$ or $M, g, s \models \psi$

$M, g, s \models \Box\, \varphi$ if and only if for all t with sRt we have $M, g, t \models \varphi$

Standard Translation for Propositional Fragment (encoded in HOL)

- "lifted" predicate $P_{i\to o}$
- $\neg_{(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda w_i\ \neg(\varphi w)$
- $\vee_{(i\to o)\to(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda\psi_{i\to o}\ \lambda w_i\ \varphi w \vee \psi w$
- $\Box_{(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda w_i\ \forall v_i\ Rwv \to \varphi v$

Validity

- $[\varphi_{i\to o}] = \forall w_i\ \varphi\ w$ resp. $[.] = \lambda\varphi_{i\to o}\ \forall w_i\ \varphi\ w$ (global)
- $[\varphi_{i\to o}]_{cw} = \varphi\ cw$ resp. $[.]_{cw} = \lambda\varphi_{i\to o}\ \ \varphi\ cw$ (local)

Kripke Style Semantics

$M, g, s \models P$ if and only if $s \in g(P)$

$M, g, s \models \neg\, \varphi$ if and only if $M, g, s \not\models \varphi$

$M, g, s \models \varphi \vee \psi$ if and only if $M, g, s \models \varphi$ or $M, g, s \models \psi$

$M, g, s \models \Box\, \varphi$ if and only if for all t with sRt we have $M, g, t \models \varphi$

Standard Translation for Propositional Fragment (encoded in HOL)

- "lifted" predicate $P_{i\to o}$
- $\neg_{(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda w_i\ \neg(\varphi w)$
- $\vee_{(i\to o)\to(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda\psi_{i\to o}\ \lambda w_i\ \varphi w \vee \psi w$
- $\Box_{(i\to o)\to(i\to o)} = \lambda\varphi_{i\to o}\ \lambda w_i\ \forall v_i\ Rwv \to \varphi v$

Validity

- $[\varphi_{i\to o}] = \forall w_i\ \varphi\ w$ resp. $[.] = \lambda\varphi_{i\to o}\ \forall w_i\ \varphi\ w$ (global)
- $[\varphi_{i\to o}]_{cw} = \varphi\ cw$ resp. $[.]_{cw} = \lambda\varphi_{i\to o}\ \ \varphi\ cw$ (local)

Modal Logic (in fact, Hybrid Logic) as a Fragment of HOL

Kripke Style Semantics

$M, g, s \models \forall x\ \varphi$ if and only if for all $d \in D$ we have $M, ([d/x]g), s \models \varphi$

Standard Translation extended for Quantifiers (and encoded in HOL)

- in HOL $\forall x_\alpha\ \varphi x$ is shorthand for $\Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \varphi x)$ —no binder needed!!!
- $\Pi_{(\alpha\to(i\to o))\to(i\to o)} = \lambda\Phi_{\alpha\to(i\to o)}\ \lambda w_i\ \Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \Phi x w)$

Example (compositionality and λ-conversion at work)

$$
\begin{aligned}
[\Box\forall x\ Px] \quad &\equiv \quad [\Box\Pi(\lambda x\ Px)] \quad \equiv \quad [\Box\Pi(\lambda x\ \lambda w\ Pxw)]\\
&\equiv \quad [\Box((\lambda\Phi\ \lambda w\ \Pi(\lambda x\ \Phi xw))(\lambda x\ \lambda w\ Pxw))]\\
&\equiv \quad [\Box(\lambda w\ \Pi(\lambda x\ (\lambda x\ \lambda w\ Pxw)xw))]\\
&\equiv \quad [\Box(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv \quad [(\lambda\varphi\ \lambda w\ \Pi(\lambda v\ Rwv \to \varphi v))(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv \quad [(\lambda w\ \Pi(\lambda v\ Rwv \to (\lambda w\ \Pi(\lambda x\ Pxw))v))]\\
&\equiv \quad [\lambda w\ \Pi(\lambda v\ Rwv \to \Pi(\lambda x\ Pxv))]\\
&\equiv \quad [\lambda w\ \forall v\ (Rwv \to \forall x\ Pxv)]\\
&\equiv \quad \forall w\ \forall v\ (Rwv \to \forall x\ Pxv)
\end{aligned}
$$

- above: possibilist quantification
- actualist quantification: $\Pi = \lambda\Phi\ \lambda w\ \Pi(\lambda x$ **existsAt x w** $\to \Phi xw)$
- also supported: local and global validity and consequence

Kripke Style Semantics

$M, g, s \models \forall x\ \varphi$ if and only if for all $d \in D$ we have $M, ([d/x]g), s \models \varphi$

Standard Translation extended for Quantifiers (and encoded in HOL)

- in HOL $\forall x_\alpha\ \varphi x$ is shorthand for $\Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \varphi x)$ —no binder needed!!!
- $\Pi_{(\alpha\to(i\to o))\to(i\to o)} = \lambda\Phi_{\alpha\to(i\to o)}\ \lambda w_i\ \Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \Phi x w)$

Example (compositionality and λ-conversion at work)

$$
\begin{aligned}
[\Box\forall x\ Px] &\equiv [\Box\Pi(\lambda x\ Px)] \equiv [\Box\Pi(\lambda x\ \lambda w\ Pxw)]\\
&\equiv [\Box((\lambda\Phi\ \lambda w\ \Pi(\lambda x\ \Phi xw))(\lambda x\ \lambda w\ Pxw))]\\
&\equiv [\Box(\lambda w\ \Pi(\lambda x\ (\lambda x\ \lambda w\ Pxw)xw))]\\
&\equiv [\Box(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv [(\lambda\varphi\ \lambda w\ \Pi(\lambda v\ Rwv \to \varphi v))(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv [(\lambda w\ \Pi(\lambda v\ Rwv \to (\lambda w\ \Pi(\lambda x\ Pxw))v))]\\
&\equiv [\lambda w\ \Pi(\lambda v\ Rwv \to \Pi(\lambda x\ Pxv))]\\
&\equiv [\lambda w\ \forall v\ (Rwv \to \forall x\ Pxv)]\\
&\equiv \forall w\ \forall v\ (Rwv \to \forall x\ Pxv)
\end{aligned}
$$

- above: possibilist quantification
- actualist quantification: $\Pi = \lambda\Phi\ \lambda w\ \Pi(\lambda x$ **existsAt x w** $\to \Phi xw)$
- also supported: local and global validity and consequence

Kripke Style Semantics

$M, g, s \models \forall x\ \varphi$ if and only if for all $d \in D$ we have $M, ([d/x]g), s \models \varphi$

Standard Translation extended for Quantifiers (and encoded in HOL)

- in HOL $\forall x_\alpha\ \varphi x$ is shorthand for $\Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \varphi x)$ —no binder needed!!!
- $\Pi_{(\alpha\to(i\to o))\to(i\to o)} = \lambda\Phi_{\alpha\to(i\to o)}\ \lambda w_i\ \Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \Phi x w)$

Example (compositionality and λ-conversion at work)

$$
\begin{aligned}
[\Box\forall x\ Px] &\equiv [\Box\Pi(\lambda x\ Px)] \equiv [\Box\Pi(\lambda x\ \lambda w\ Pxw)]\\
&\equiv [\Box((\lambda\Phi\ \lambda w\ \Pi(\lambda x\ \Phi xw))(\lambda x\ \lambda w\ Pxw))]\\
&\equiv [\Box(\lambda w\ \Pi(\lambda x\ (\lambda x\ \lambda w\ Pxw)xw))]\\
&\equiv [\Box(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv [(\lambda\varphi\ \lambda w\ \Pi(\lambda v\ Rwv \to \varphi v))(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv [(\lambda w\ \Pi(\lambda v\ Rwv \to (\lambda w\ \Pi(\lambda x\ Pxw))v))]\\
&\equiv [\lambda w\ \Pi(\lambda v\ Rwv \to \Pi(\lambda x\ Pxv))]\\
&\equiv [\lambda w\ \forall v\ (Rwv \to \forall x\ Pxv)]\\
&\equiv \forall w\ \forall v\ (Rwv \to \forall x\ Pxv)
\end{aligned}
$$

- above: possibilist quantification
- actualist quantification: $\Pi = \lambda\Phi\ \lambda w\ \Pi(\lambda x$ **existsAt x w** $\to \Phi xw)$
- also supported: local and global validity and consequence

Kripke Style Semantics

$M, g, s \models \forall x\ \varphi$ if and only if for all $d \in D$ we have $M, ([d/x]g), s \models \varphi$

Standard Translation extended for Quantifiers (and encoded in HOL)

- in HOL $\forall x_\alpha\ \varphi x$ is shorthand for $\Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \varphi x)$ —no binder needed!!!
- $\Pi_{(\alpha\to(i\to o))\to(i\to o)} = \lambda\Phi_{\alpha\to(i\to o)}\ \lambda w_i\ \Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \Phi x w)$

Example (compositionality and λ-conversion at work)

$$
\begin{aligned}
[\Box\forall x\ Px] \quad &\equiv \quad [\Box\Pi(\lambda x\ Px)] \quad \equiv \quad [\Box\Pi(\lambda x\ \lambda w\ Pxw)]\\
&\equiv \quad [\Box((\lambda\Phi\ \lambda w\ \Pi(\lambda x\ \Phi xw))(\lambda x\ \lambda w\ Pxw))]\\
&\equiv \quad [\Box(\lambda w\ \Pi(\lambda x\ (\lambda x\ \lambda w\ Pxw)xw))]\\
&\equiv \quad [\Box(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv \quad [(\lambda\varphi\ \lambda w\ \Pi(\lambda v\ Rwv \to \varphi v))(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv \quad [(\lambda w\ \Pi(\lambda v\ Rwv \to (\lambda w\ \Pi(\lambda x\ Pxw))v))]\\
&\equiv \quad [\lambda w\ \Pi(\lambda v\ Rwv \to \Pi(\lambda x\ Pxv))]\\
&\equiv \quad [\lambda w\ \forall v\ (Rwv \to \forall x\ Pxv)]\\
&\equiv \quad \forall w\ \forall v\ (Rwv \to \forall x\ Pxv)
\end{aligned}
$$

- above: possibilist quantification
- actualist quantification: $\Pi = \lambda\Phi\ \lambda w\ \Pi(\lambda x$ **existsAt x w** $\to \Phi xw)$
- also supported: local and global validity and consequence

Kripke Style Semantics

$M, g, s \models \forall x\ \varphi$ if and only if for all $d \in D$ we have $M, ([d/x]g), s \models \varphi$

Standard Translation extended for Quantifiers (and encoded in HOL)

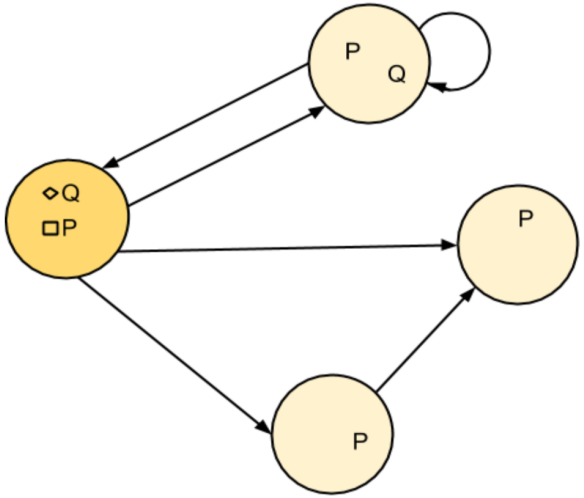
- in HOL $\forall x_\alpha\ \varphi x$ is shorthand for $\Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \varphi x)$ —no binder needed!!!
- $\Pi_{(\alpha\to(i\to o))\to(i\to o)} = \lambda\Phi_{\alpha\to(i\to o)}\ \lambda w_i\ \Pi_{(\alpha\to o)\to o}(\lambda x_\alpha\ \Phi x w)$

Example (compositionality and λ-conversion at work)

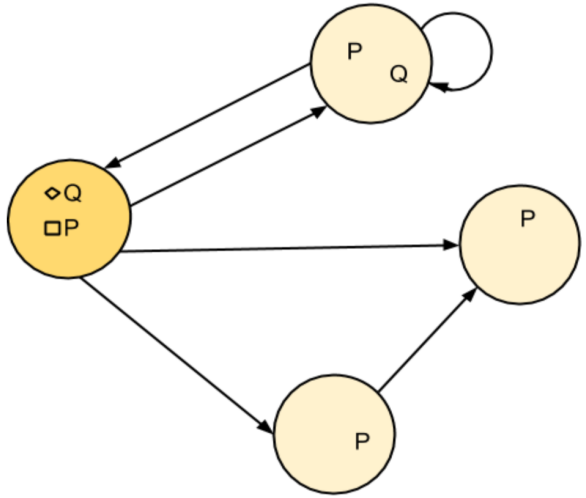
$$
\begin{aligned}
[\Box\forall x\ Px] &\equiv [\Box\Pi(\lambda x\ Px)] \equiv [\Box\Pi(\lambda x\ \lambda w\ Pxw)]\\
&\equiv [\Box((\lambda\Phi\ \lambda w\ \Pi(\lambda x\ \Phi xw))(\lambda x\ \lambda w\ Pxw))]\\
&\equiv [\Box(\lambda w\ \Pi(\lambda x\ (\lambda x\ \lambda w\ Pxw)xw))]\\
&\equiv [\Box(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv [(\lambda\varphi\ \lambda w\ \Pi(\lambda v\ Rwv \to \varphi v))(\lambda w\ \Pi(\lambda x\ Pxw))]\\
&\equiv [(\lambda w\ \Pi(\lambda v\ Rwv \to (\lambda w\ \Pi(\lambda x\ Pxw))v))]\\
&\equiv [\lambda w\ \Pi(\lambda v\ Rwv \to \Pi(\lambda x\ Pxv))]\\
&\equiv [\lambda w\ \forall v\ (Rwv \to \forall x\ Pxv)]\\
&\equiv \forall w\ \forall v\ (Rwv \to \forall x\ Pxv)
\end{aligned}
$$

- above: possibilist quantification
- actualist quantification: $\Pi = \lambda\Phi\ \lambda w\ \Pi(\lambda x$ **existsAt x w** $\to \Phi xw)$
- also supported: local and global validity and consequence

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

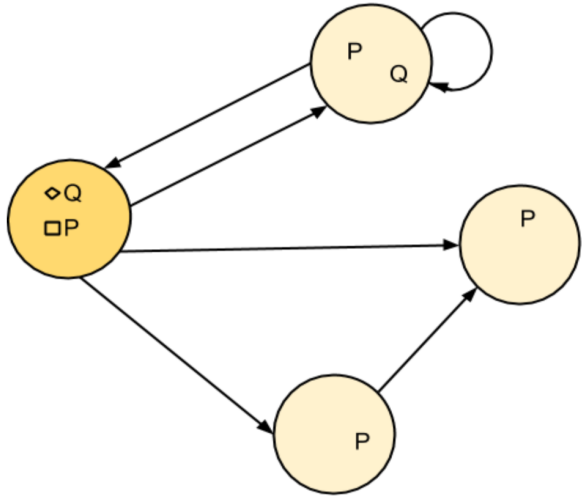


Properties of □ and ◇ correlated to structure of transition system between worlds



- 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 $\forall P.\Box P \rightarrow \Box\Box P$

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

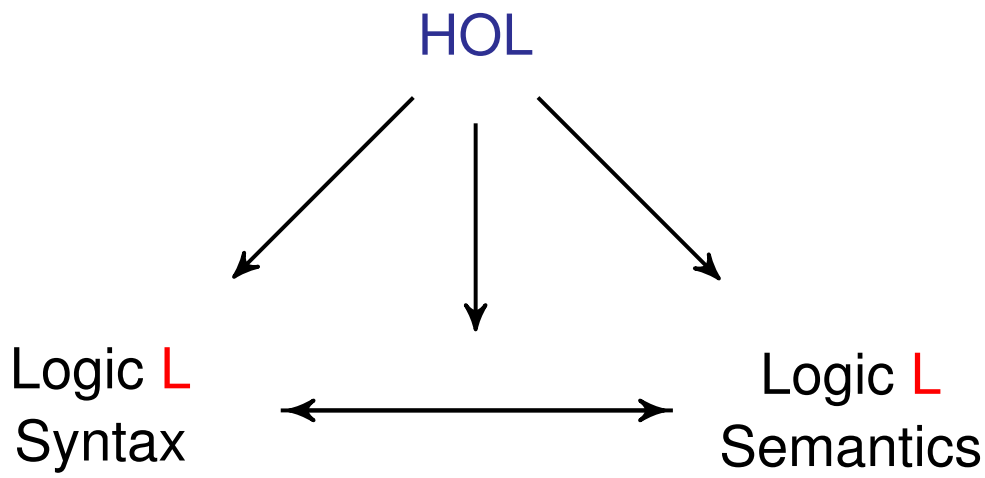


- Logic K: — (no restrictions, any structure)
- Logic M: reflexiv transition relation, $\forall P.\Box P \rightarrow P$
- Logic KB: symmetric transition relation, $\forall P.P \rightarrow \Box\Diamond P$
- Logic S5: equivelance relation as transition system, add $\forall P.\Box P \rightarrow \Box\Box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

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

  alternatively: $\forall P.\neg(\Box P \wedge \Box\neg P)$ or $\forall x.\exists y.xRy$

Flexible Reasoning Technology: SSEs in 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, Deontic Logics, ...

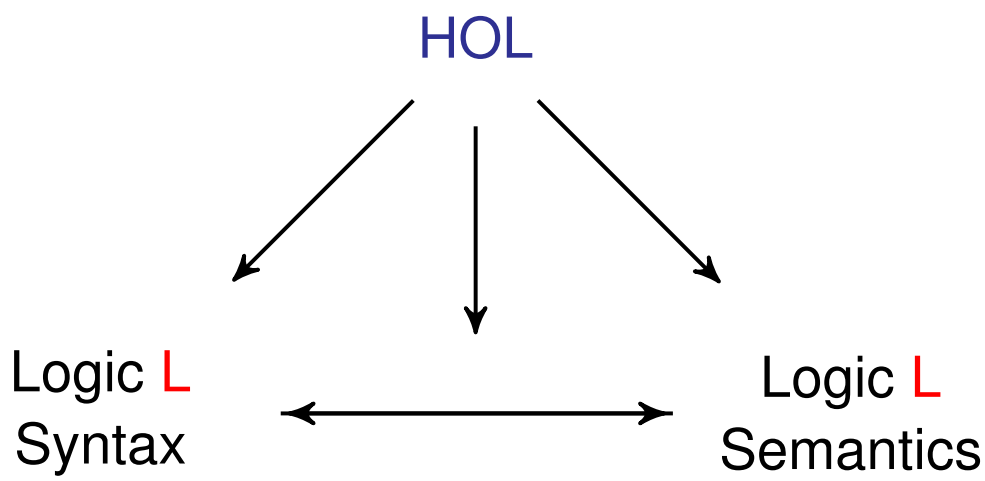
Embedding works also for quantifiers (first-order & higher-order)

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, ...

Flexible Reasoning Technology: SSEs in 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, Deontic Logics, ...

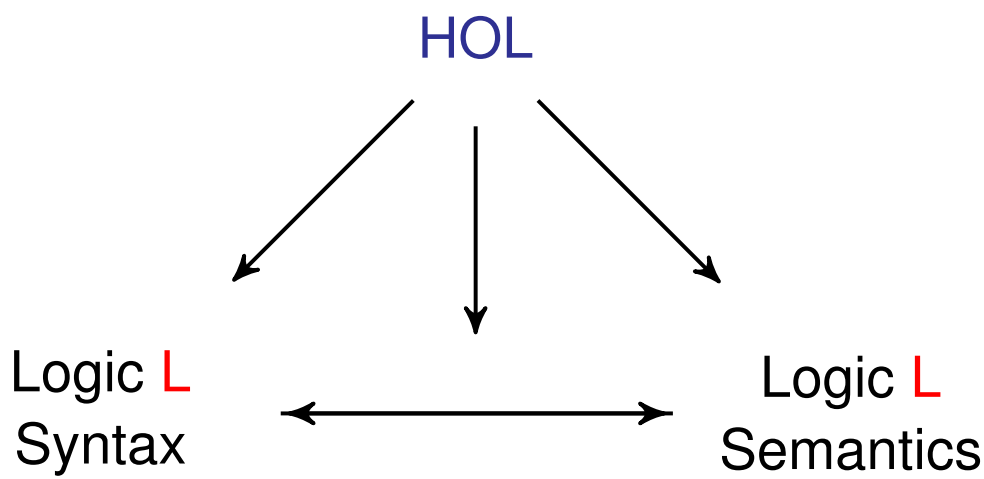
Embedding works also for quantifiers (first-order & higher-order)

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, ...

Flexible Reasoning Technology: SSEs in 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, Deontic Logics, ...

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

HOL provers become universal logic reasoning engines!

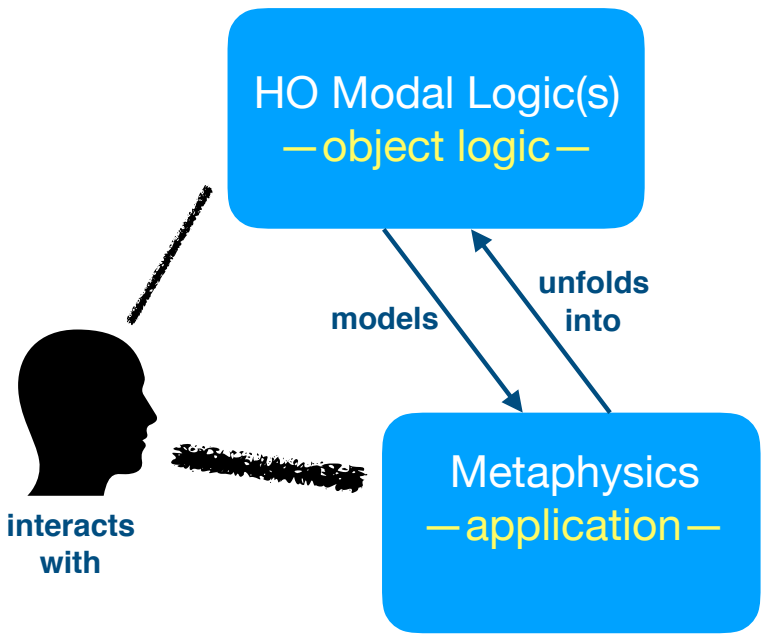
interactive:	Isabelle/HOL, PVS, HOL4, Hol Light, Coq/HOL, ...
automated:	Leo-III, LEO-II, Satallax, TPS, Nitpick, Isabelle/HOL, ...

Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]

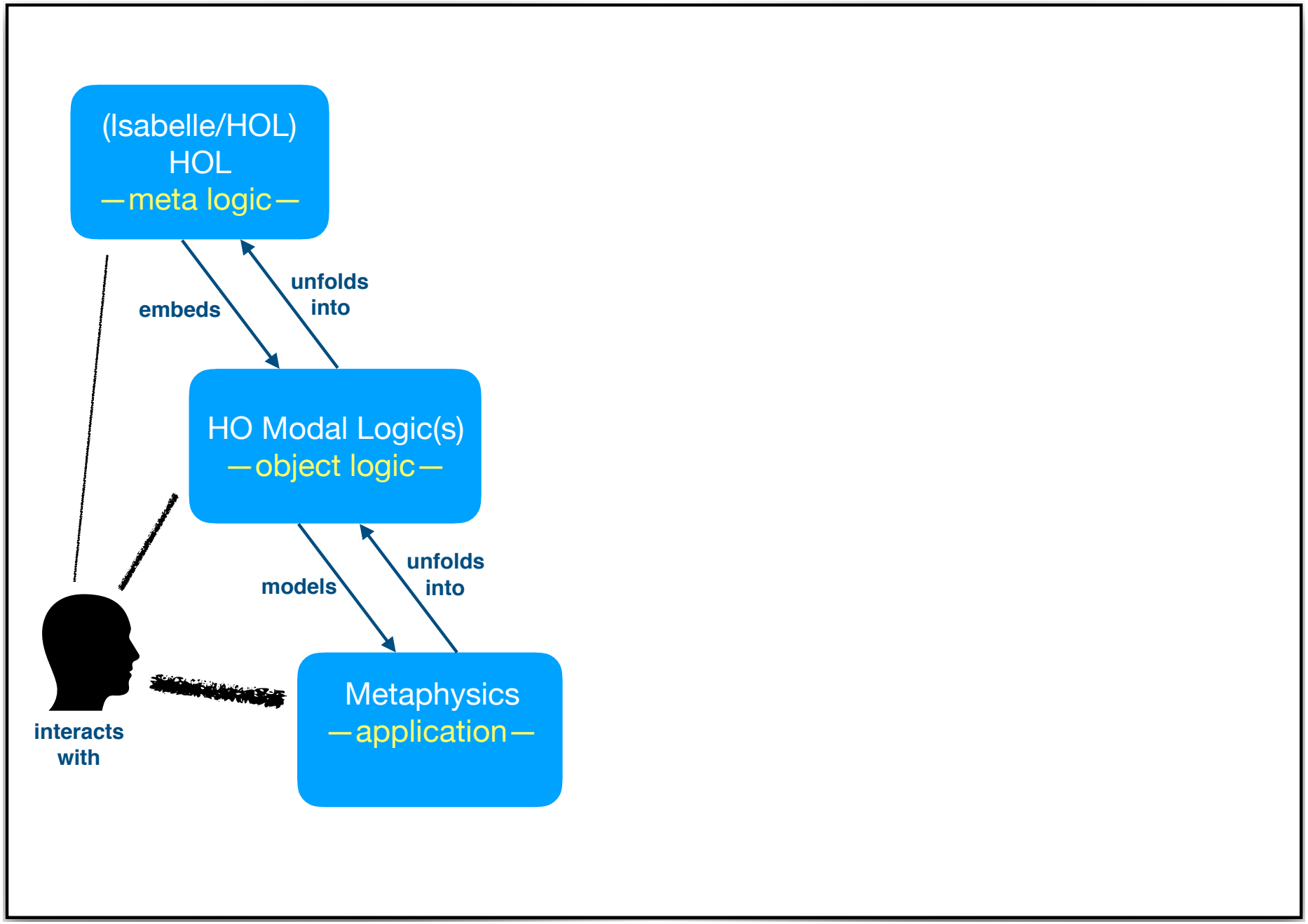
interacts with

Metaphysics
—application—

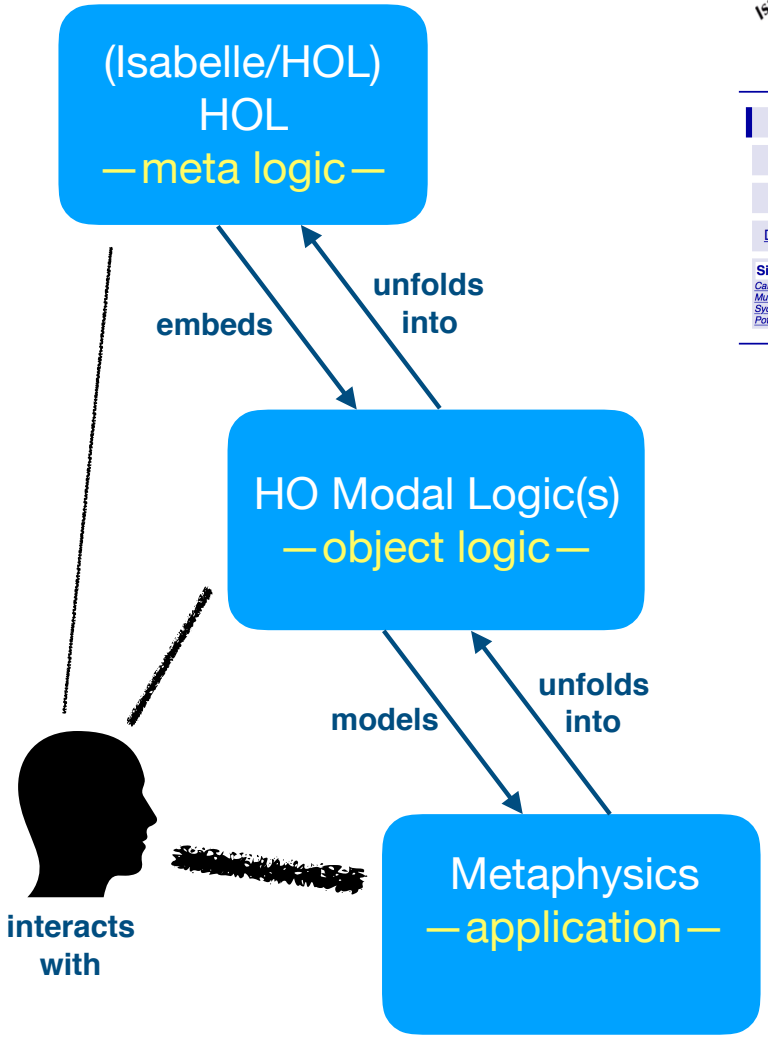
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]

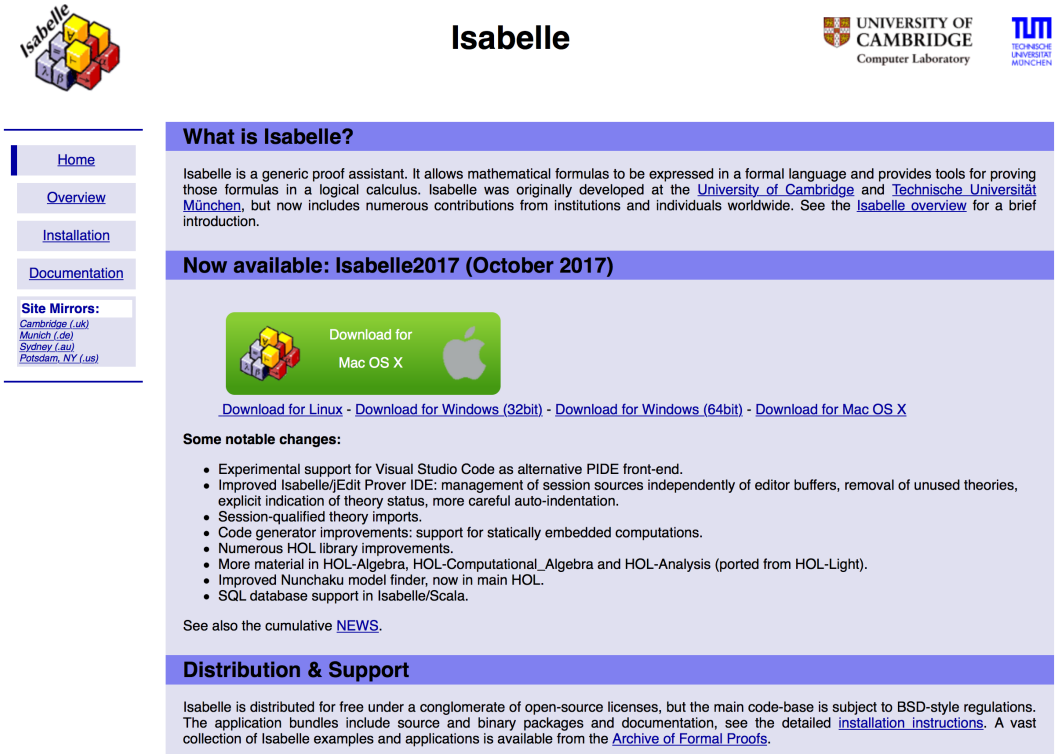


Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]

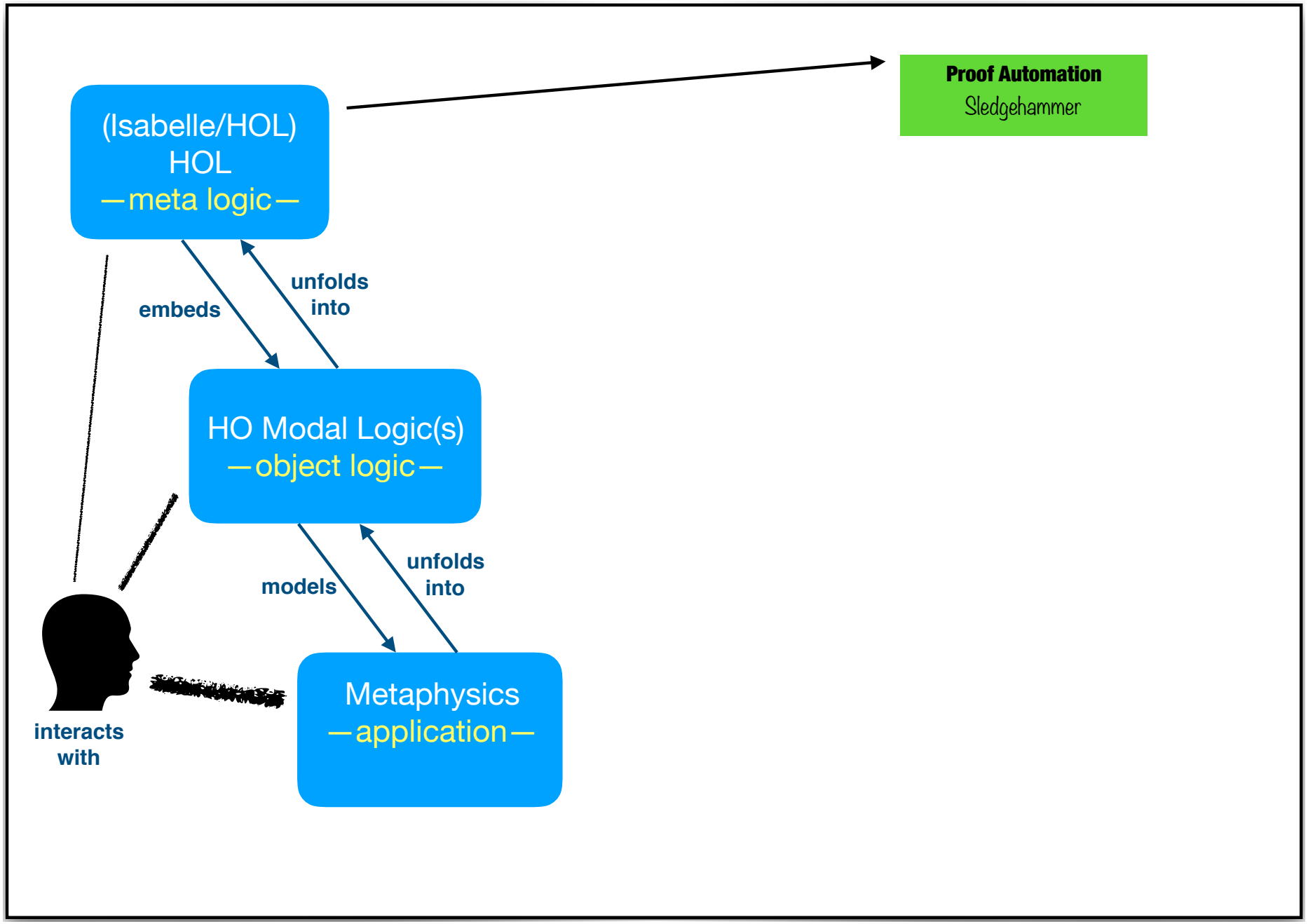


Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]

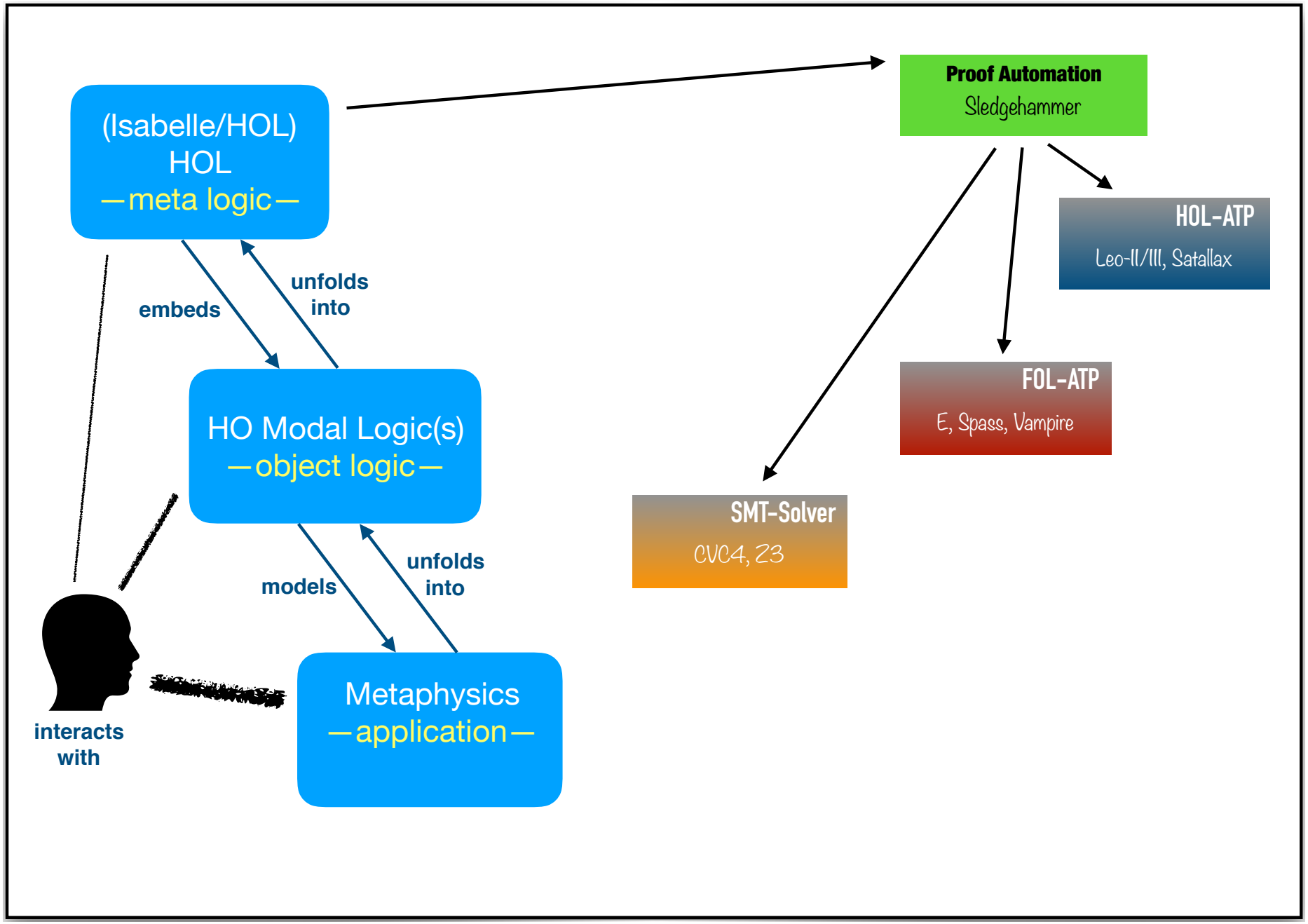




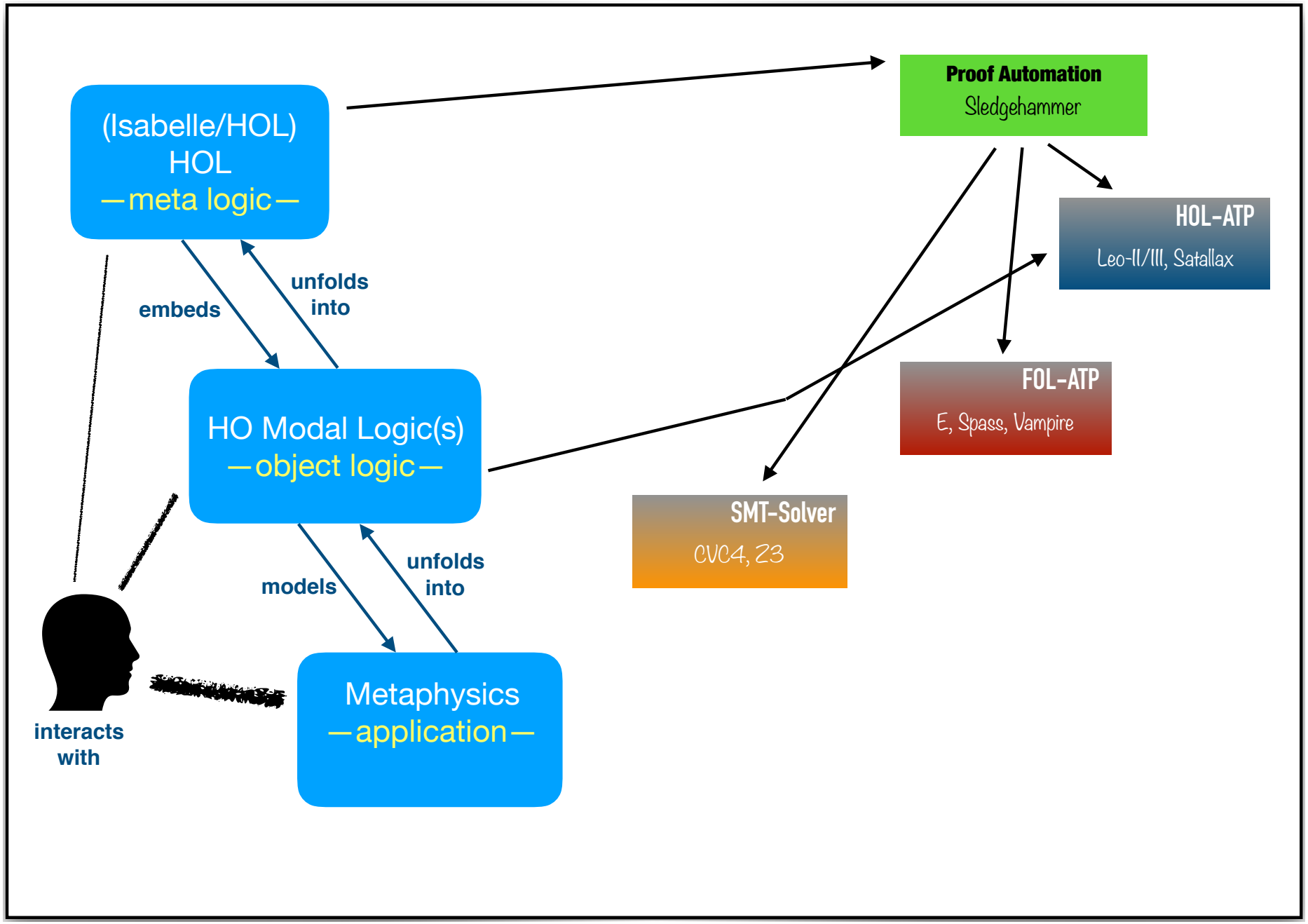
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



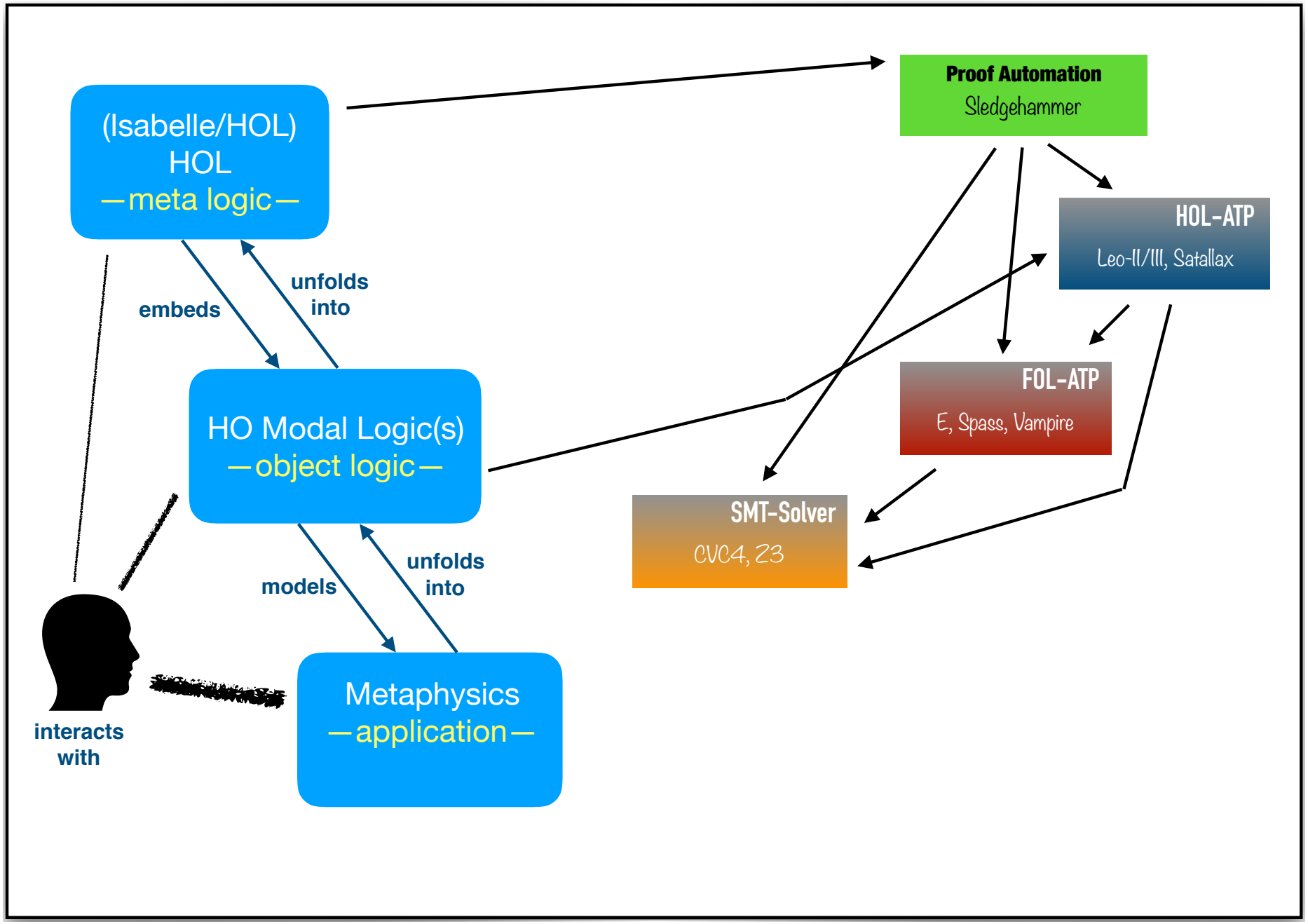
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



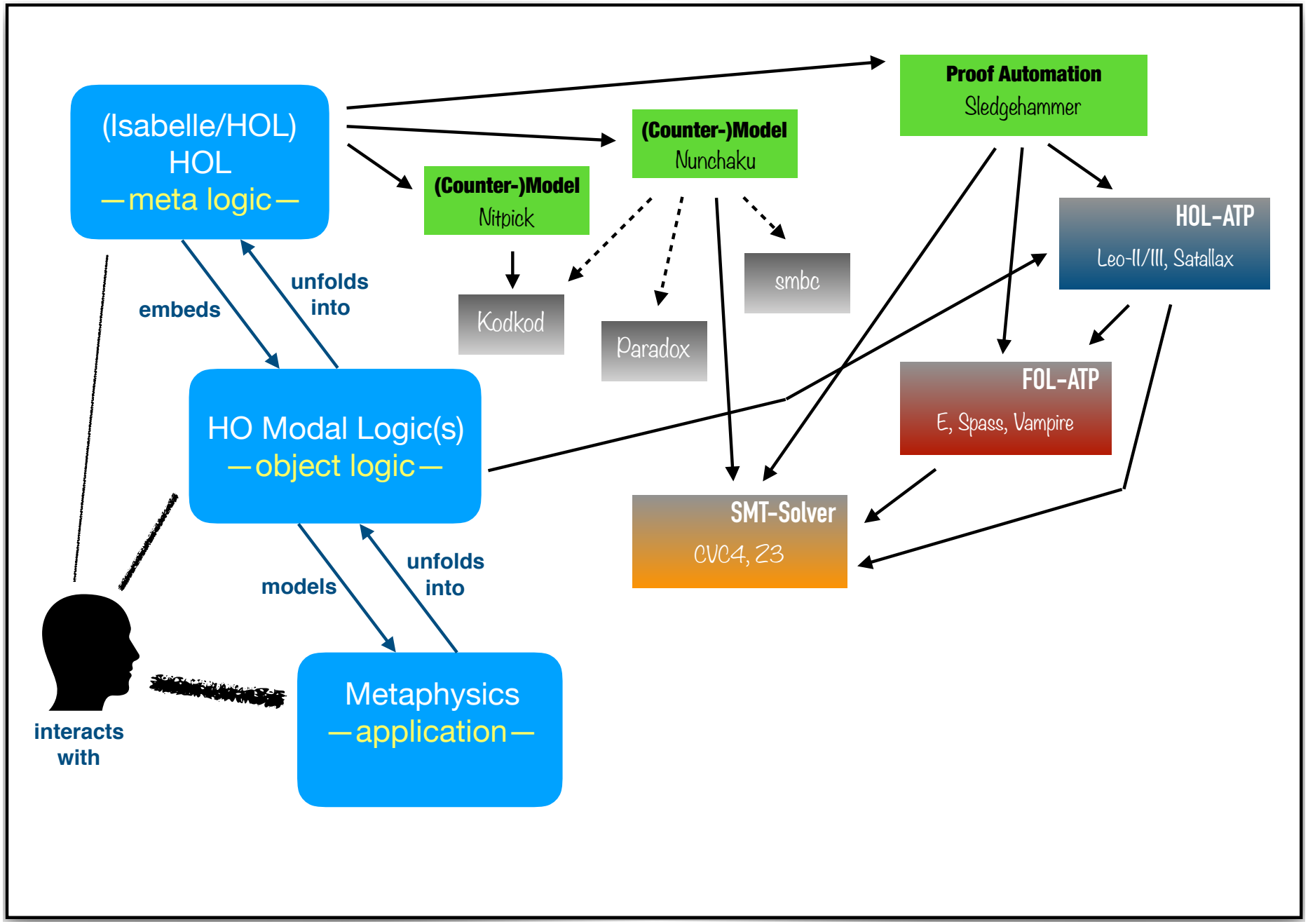
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



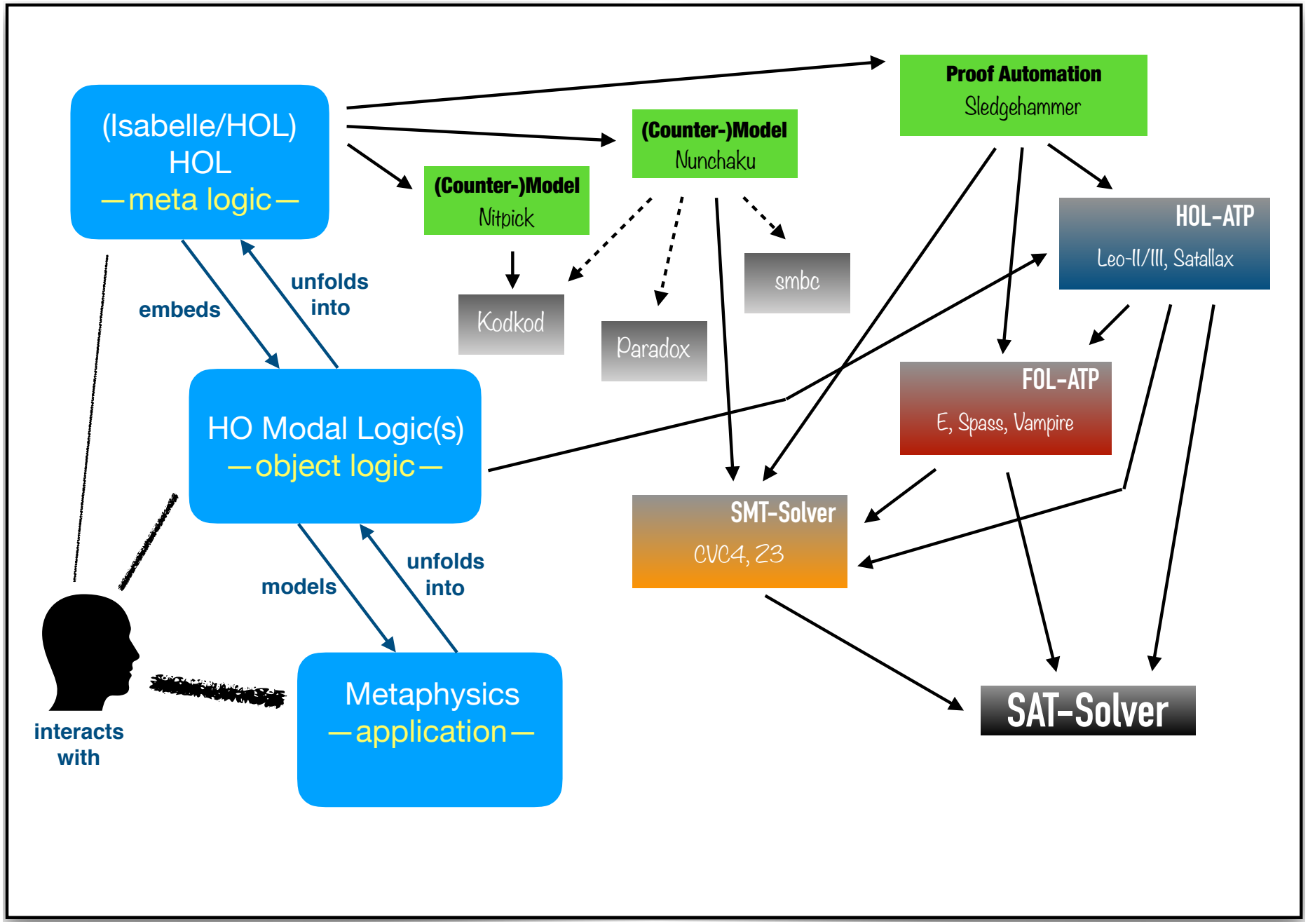
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



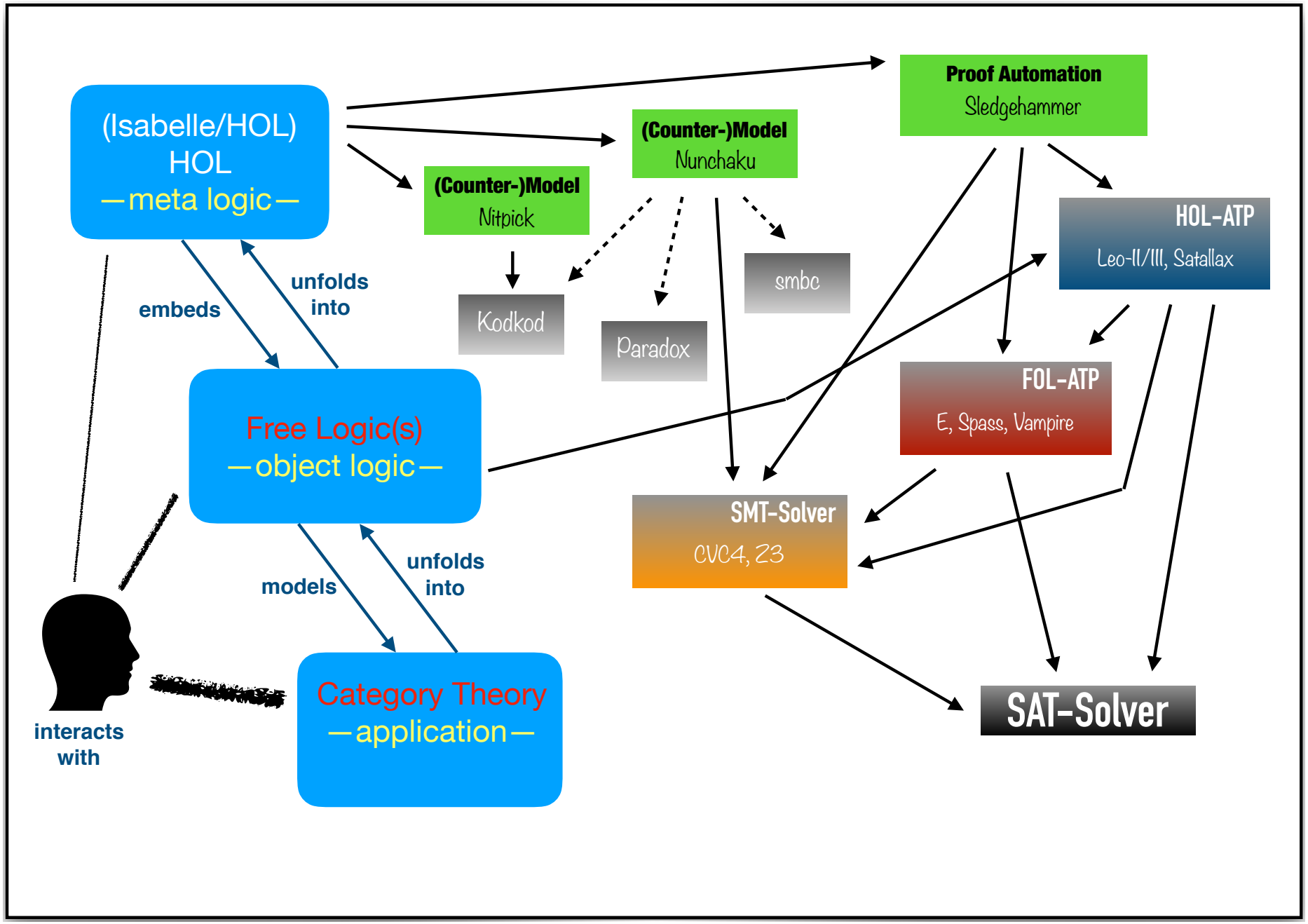
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



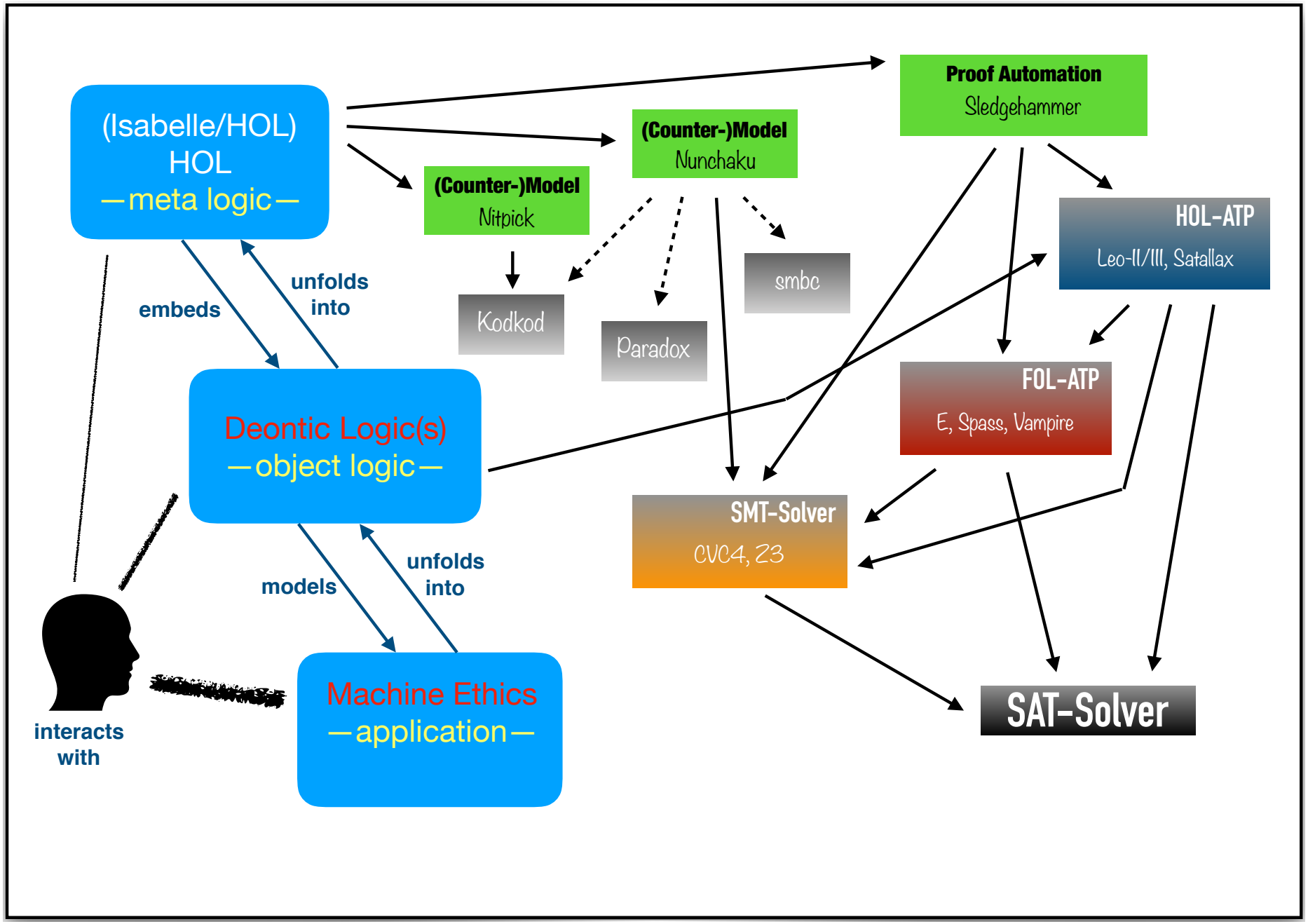
Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



Flexible Reasoning Technology: SSEs in HOL [SCP, 2019], [OpenPhilos., 2019]



Flexible Reasoning Technology: SSEs in HOL

Ed Zalta (Stanford)

Principia Logico-Metaphysica

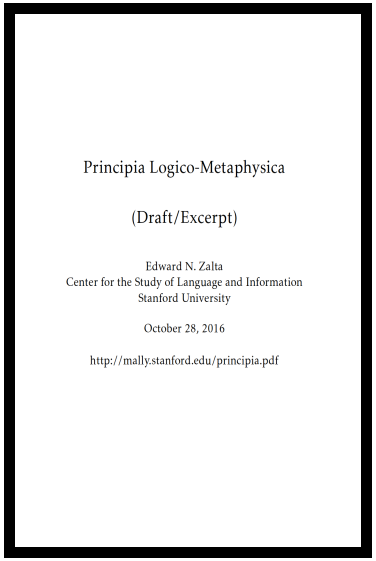
Hyperintensional higher-order modal logic

Inconsistency/Paradox detected & fixed

[Open Philosophy, in print] [RSL, in print]

Daniel Kirchner
(Mathematics, FU Berlin)

Flexible Reasoning Technology: SSEs in HOL


Principia Logico-Metaphysica

(Draft/Excerpt)

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

October 28, 2016

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

Ed Zalta (Stanford)

Principia Logico-Metaphysica

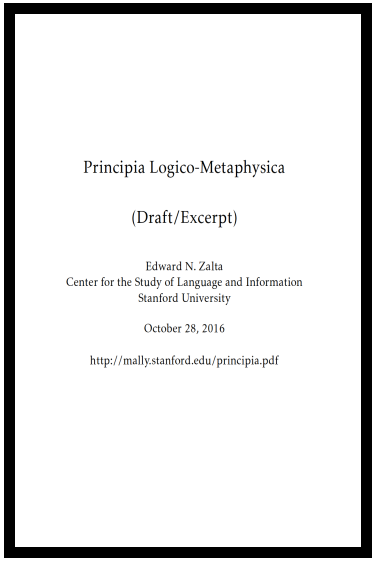
Hyperintensional higher-order modal logic

Inconsistency/Paradox detected & fixed

[Open Philosophy, in print] [RSL, in print]


Daniel Kirchner
(Mathematics, FU Berlin)

Flexible Reasoning Technology: SSEs in HOL


Principia Logico-Metaphysica

(Draft/Excerpt)

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

October 28, 2016

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

Ed Zalta (Stanford)

Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected & fixed

[Open Philosophy, in print] [RSL, in print]

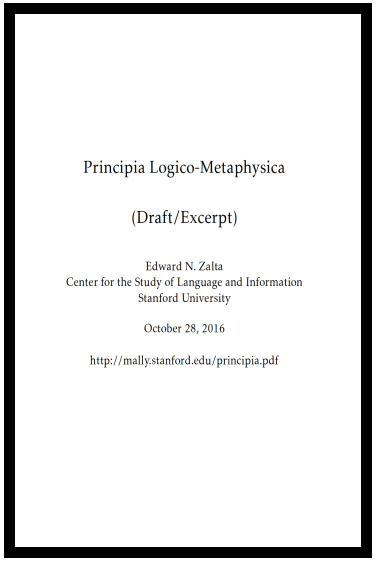

Daniel Kirchner
(Mathematics, FU Berlin)

Kirchner Paradox

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

Computational Metaphysics par excellence!!!

Flexible Reasoning Technology: SSEs in HOL



Ed Zalta (Stanford)

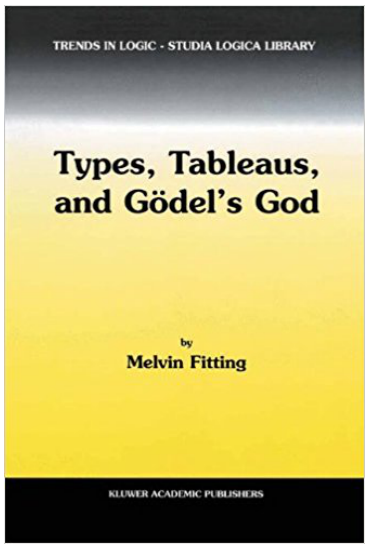
Principia Logico-Metaphysica

Hyperintensional higher-order modal logic

Inconsistency/Paradox detected & fixed

[Open Philosophy, in print] [RSL, in print]

Daniel Kirchner
(Mathematics, FU Berlin)



M. Fitting (New York)

Philosophy/Metaphysics

Ontological Argument
(avoids modal collapse)

Intensional higher-order modal logic

Verified (main chapters)

[Archive of Formal Proofs, 2017]

David Fuenmayor
(Philosophy, FU Berlin)

Flexible Reasoning Technology: SSEs in HOL

Ontological Proofs Today,
edited by Miroslaw Szatkowski,
Ontos Verlag, 2012

A New Modal Version of the Ontological Argument

E. J. Lowe

The original version of the ontological argument for the existence of God is due to St Anselm (see Charlesworth [1]). He argued that God is, by definition, a being than which none greater can be conceived; that such a being exists at least in the mind of the conceiver, that is, in conception; that it is greater to exist in reality than to exist merely in conception; and hence that such a being does exist in reality. This argument speaks explicitly only of God's *existence*, not of his *necessary* existence. Modal versions of the ontological argument speak of the latter. Typically, they run roughly as follows. God is, by definition, a *necessary* being – one that *necessarily* exists; but, given that such a being is *possible*, it exists at least in *some* possible world, w; however, since it is thus true in w that this necessary being exists, it is true in w that this being exists in *every* possible world, including the actual world; hence, this being exists also in actuality (see Plantinga [7], ch. 10). There are many objections that can be and have been raised against both modal and non-modal versions of the ontological argument. Rather than try to rebut those objections or refine existing versions of the argument,[1] what I shall do in this chapter is to develop a new kind of modal ontological argument for the existence of God – or, more precisely and somewhat less ambitiously, for the existence of a *necessary concrete* being. I shall explain in due course why, although this kind of argument differs from more familiar variants of the modal

[1] I try to do both in Lowe [4], where I also say more about the history of the ontological argument.

179

E. J. Lowe's (Durham)

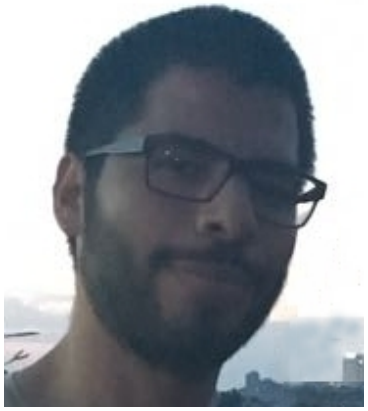
Philosophy/Metaphysics

E. J. Lowe's Modal Ontological Argument

Higher-order modal logic

Verified (minor corrections)

[Journal of Applied Logics, 2018]



David Fuenmayor
(Philosophy, FU Berlin)

Flexible Reasoning Technology: SSEs in HOL

Ontological Proofs Today, edited by Miroslaw Szatkowski, Ontos Verlag, 2012

A New Modal Version of the Ontological Argument

E. J. Lowe

The original version of the ontological argument for the existence of God is due to St Anselm (see Charlesworth [1]). He argued that God is, by definition, a being than which none greater can be conceived; that such a being exists at least in the mind of the conceiver, that is, in conception; that it is greater to exist in reality than to exist merely in conception; and hence that such a being does exist in reality. This argument speaks explicitly only of God's *existence*, not of his *necessary* existence. Modal versions of the ontological argument speak of the latter. Typically, they run roughly as follows. God is, by definition, a *necessary* being – one that *necessarily* exists; but, given that such a being is *possible*, it exists at least in *some* possible world, w; however, since it is thus true in w that this necessary being exists, it is true in w that this being exists in *every* possible world, including the actual world; hence, this being exists also in actuality (see Plantinga [7], ch. 10). There are many objections that can be and have been raised against both modal and non-modal versions of the ontological argument. Rather than try to rebut those objections or refine existing versions of the argument,[1] what I shall do in this chapter is to develop a new kind of modal ontological argument for the existence of God – or, more precisely and somewhat less ambitiously, for the existence of a *necessary concrete* being. I shall explain in due course why, although this kind of argument differs from more familiar variants of the modal

[1] I try to do both in Lowe [4], where I also say more about the history of the ontological argument.

179

E. J. Lowe's (Durham)

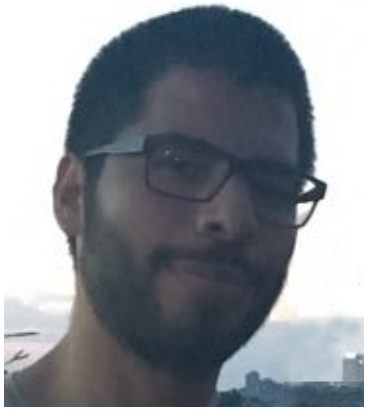
Philosophy/Metaphysics

E. J. Lowe's Modal Ontological Argument

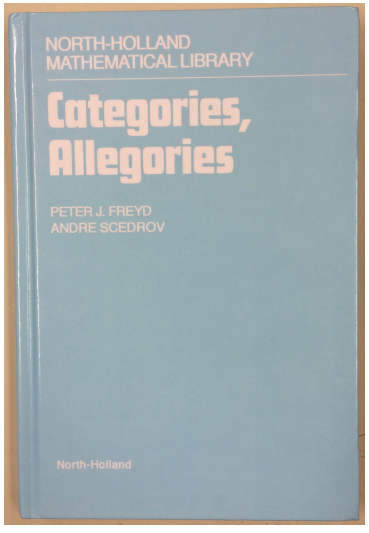
Higher-order modal logic

Verified (minor corrections)

[Journal of Applied Logics, 2018]



David Fuenmayor
(Philosophy, FU Berlin)



P. Freyd & A. Scedrov

Category Theory

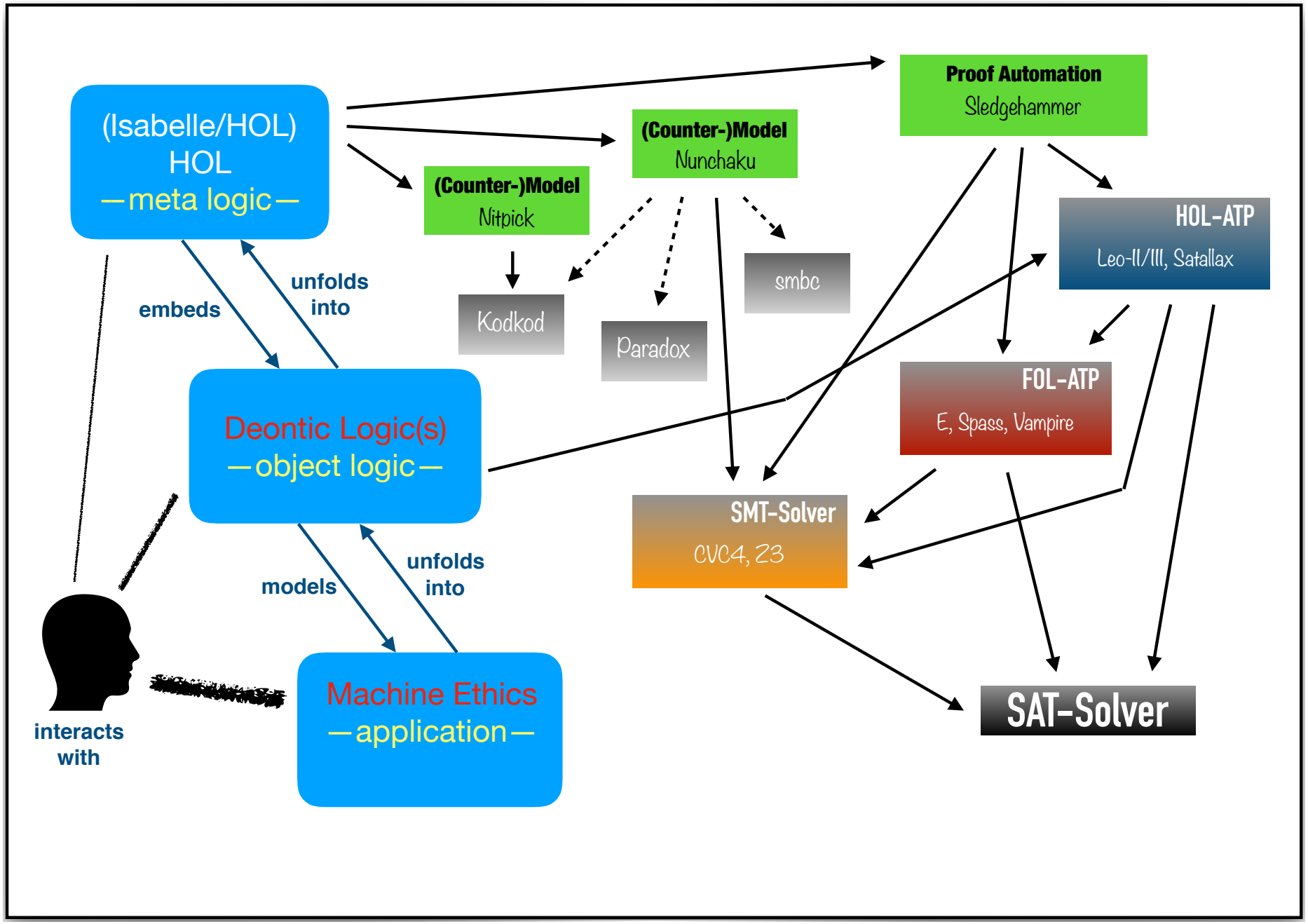
Free first-order logic

(Constricted) Inconsistency detected & fixed

[JAR, 2019]

D. Scott
(UC Berkeley)

Flexible Reasoning Technology: SSEs in HOL



Which Reasoning Formalisms?

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

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

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

Which Reasoning Formalisms?

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

(Leibniz, 1677)

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

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

Standard CTD structure (Chisholm)

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

Danger: Paradox/inconsistency — ex falso quodlibet!

Which Reasoning Formalisms?

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

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

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

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

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

Danger: Paradox/inconsistency — ex falso quodlibet!

Which Reasoning Formalisms?

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

(Leibniz, 1677)

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

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

L. van der Torre

X. Parent

A. Farjami

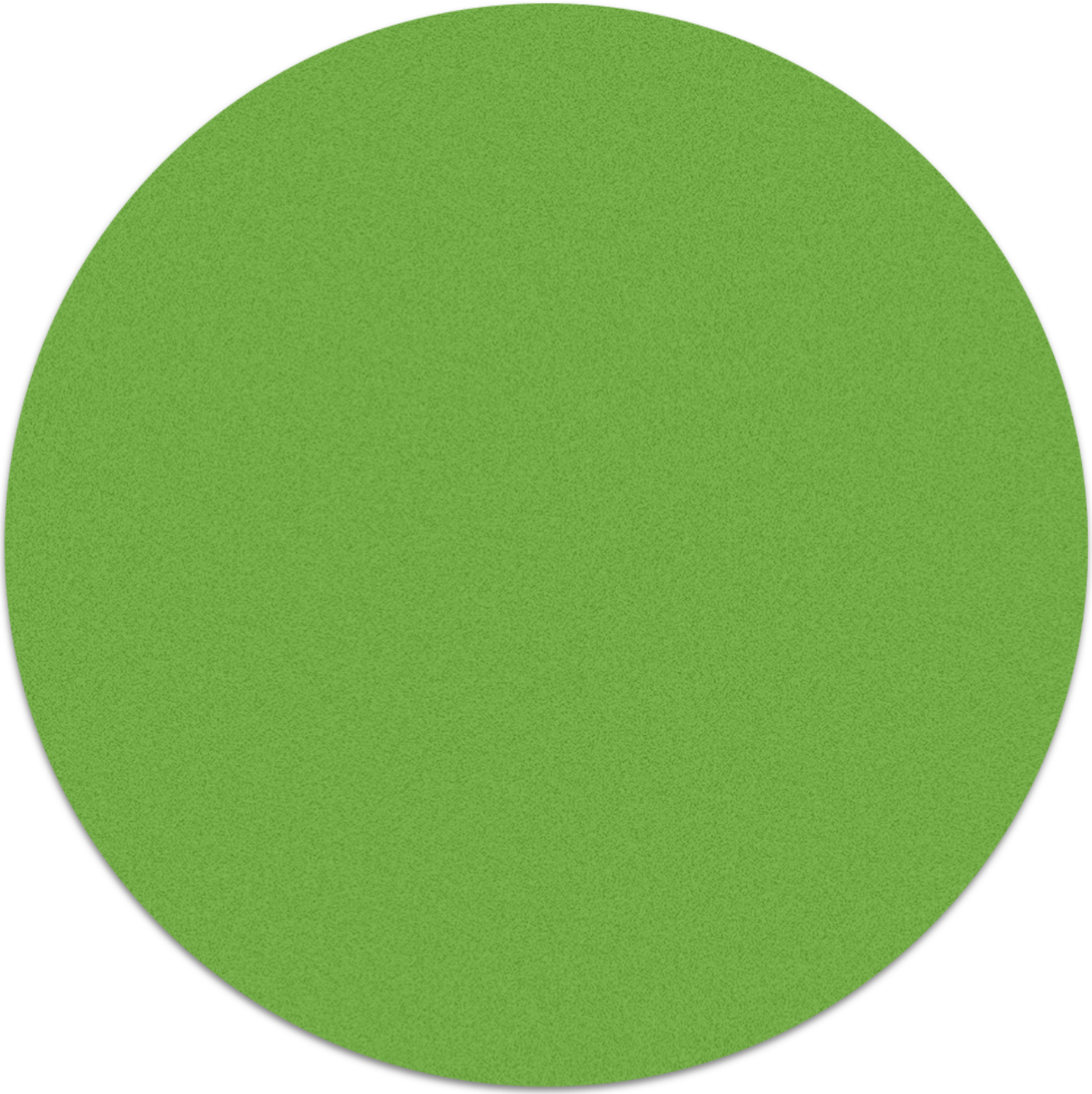
Deontic Logic

- Reasoning about obligations and permissions
- Two groups of approaches:
  - — Possible worlds
    - standard deontic logic — **CTD: no**
    - dyadic deontic logic — **CTD: yes**
  - — Norm-based semantics
    - input/output logic — **CTD: yes**

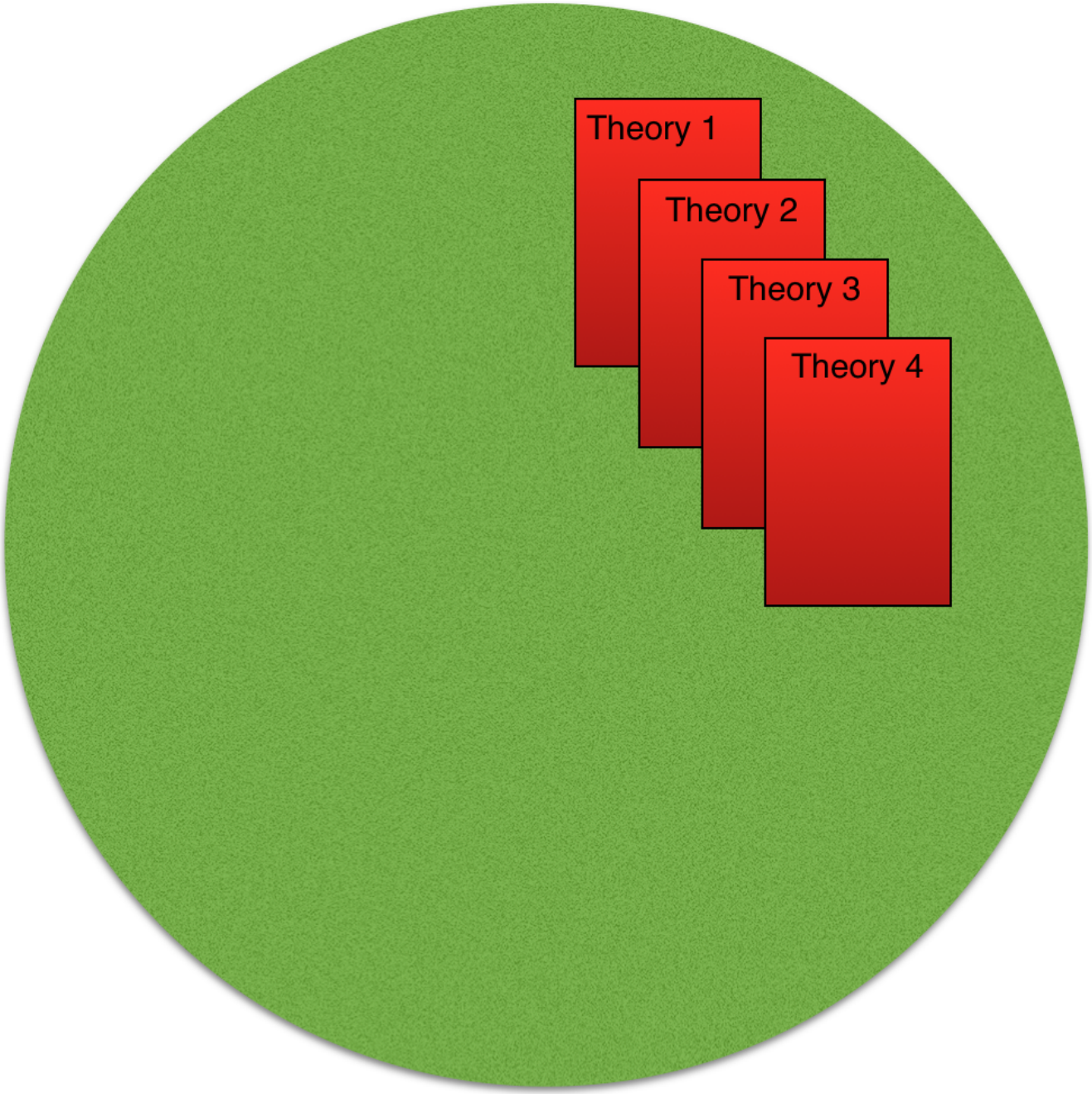
Further interests and challenges

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

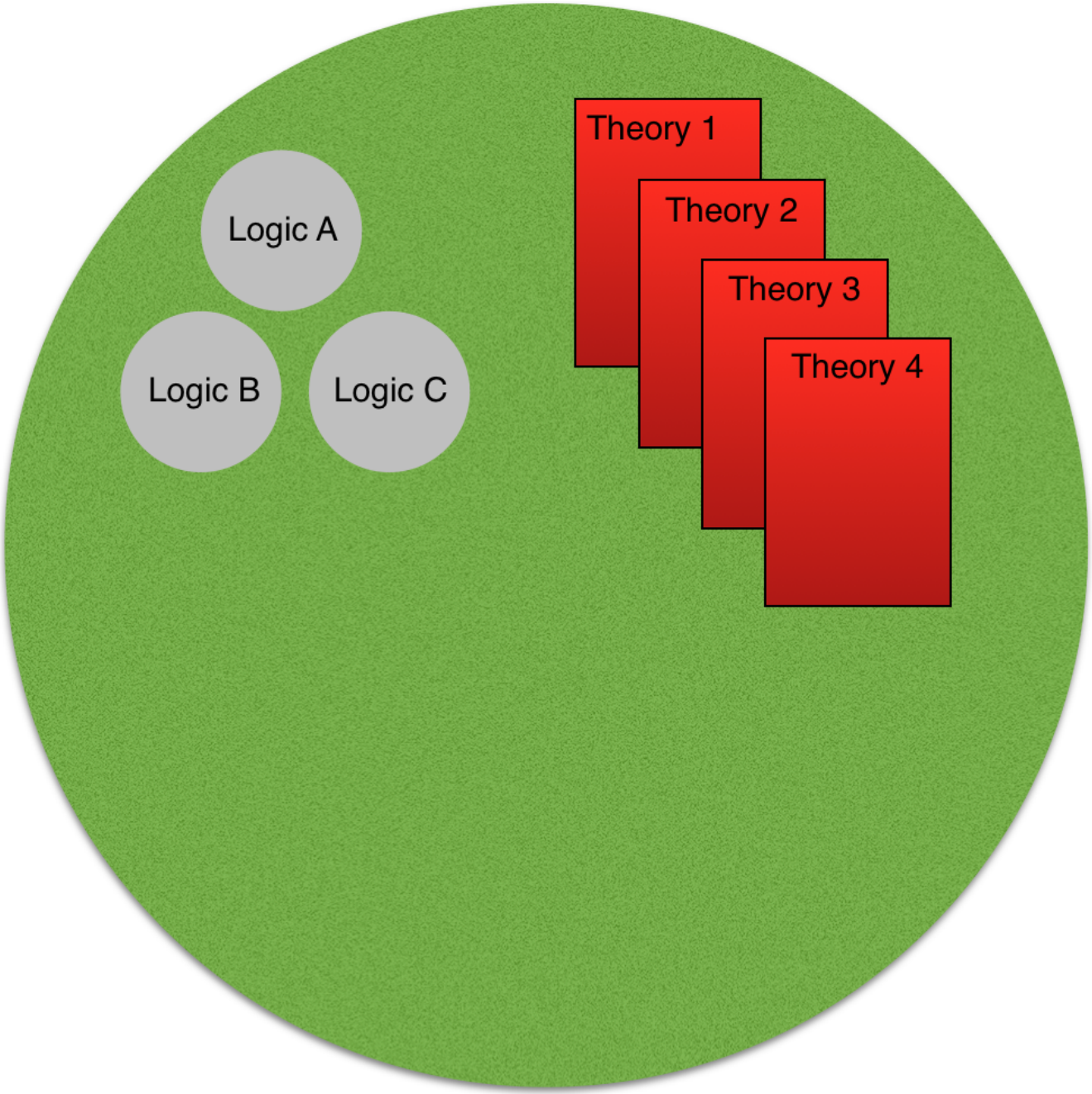
Normative Reasoning Experimentation Platform



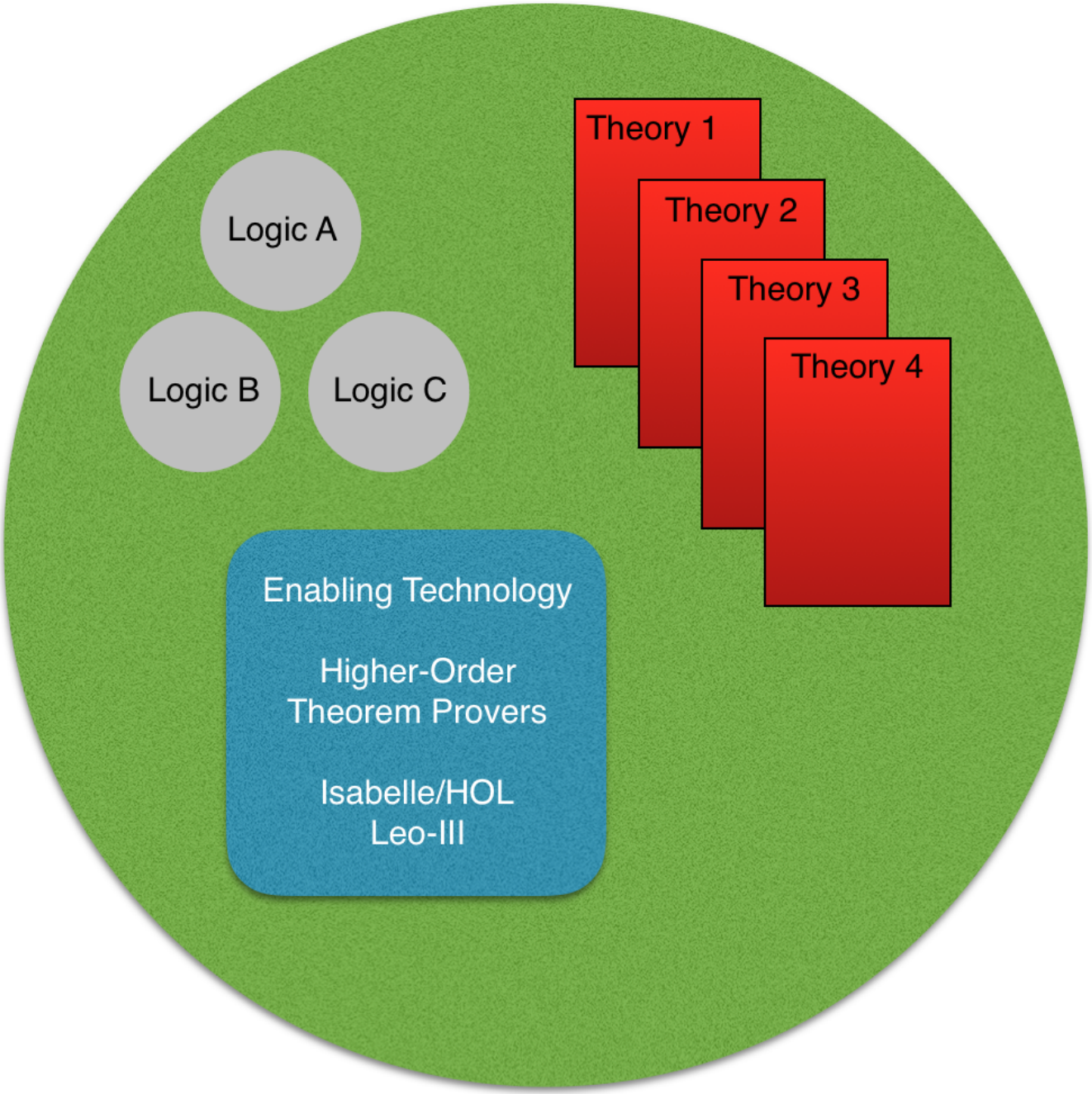
Normative Reasoning Experimentation Platform



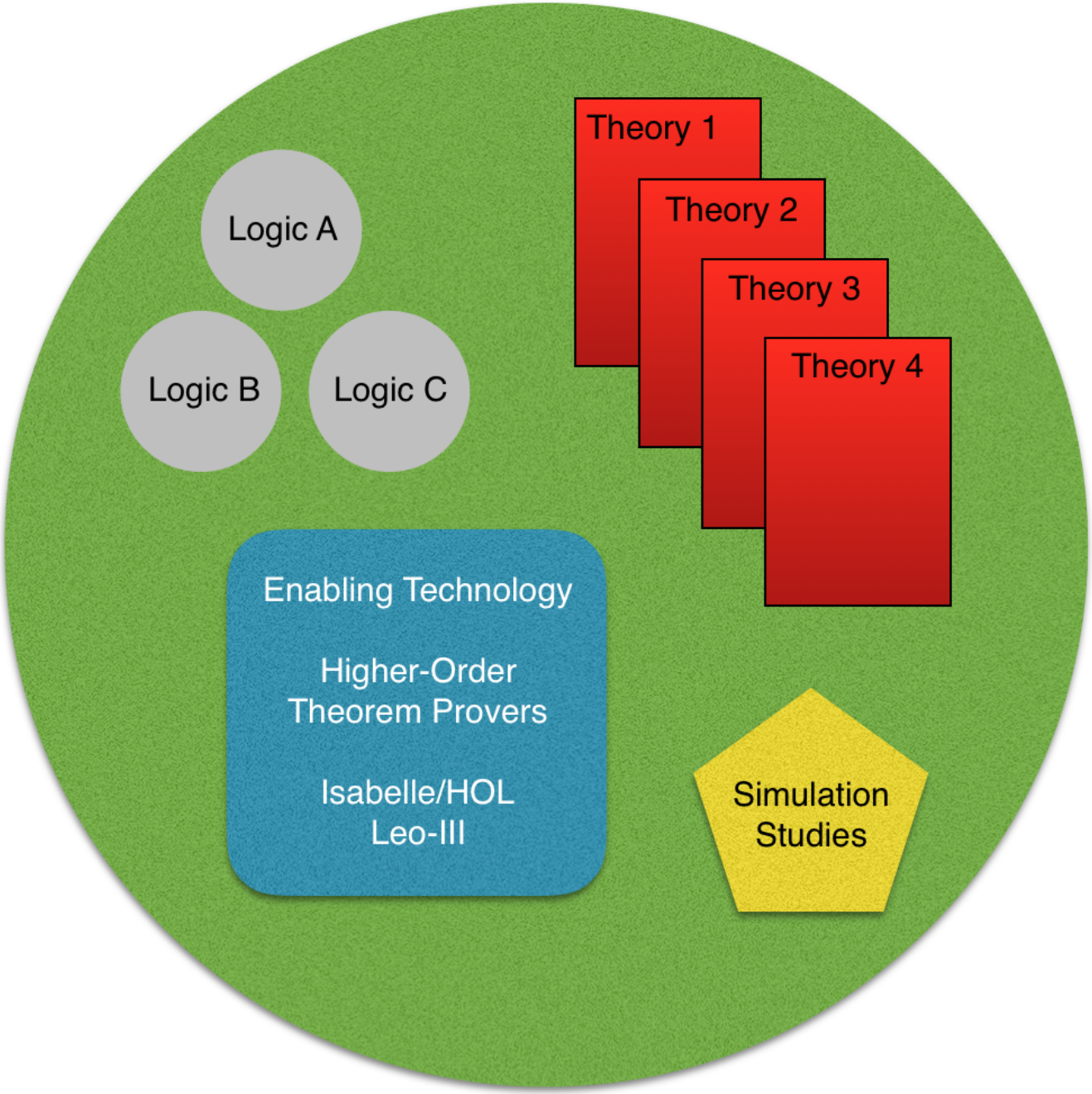
Normative Reasoning Experimentation Platform



Normative Reasoning Experimentation Platform

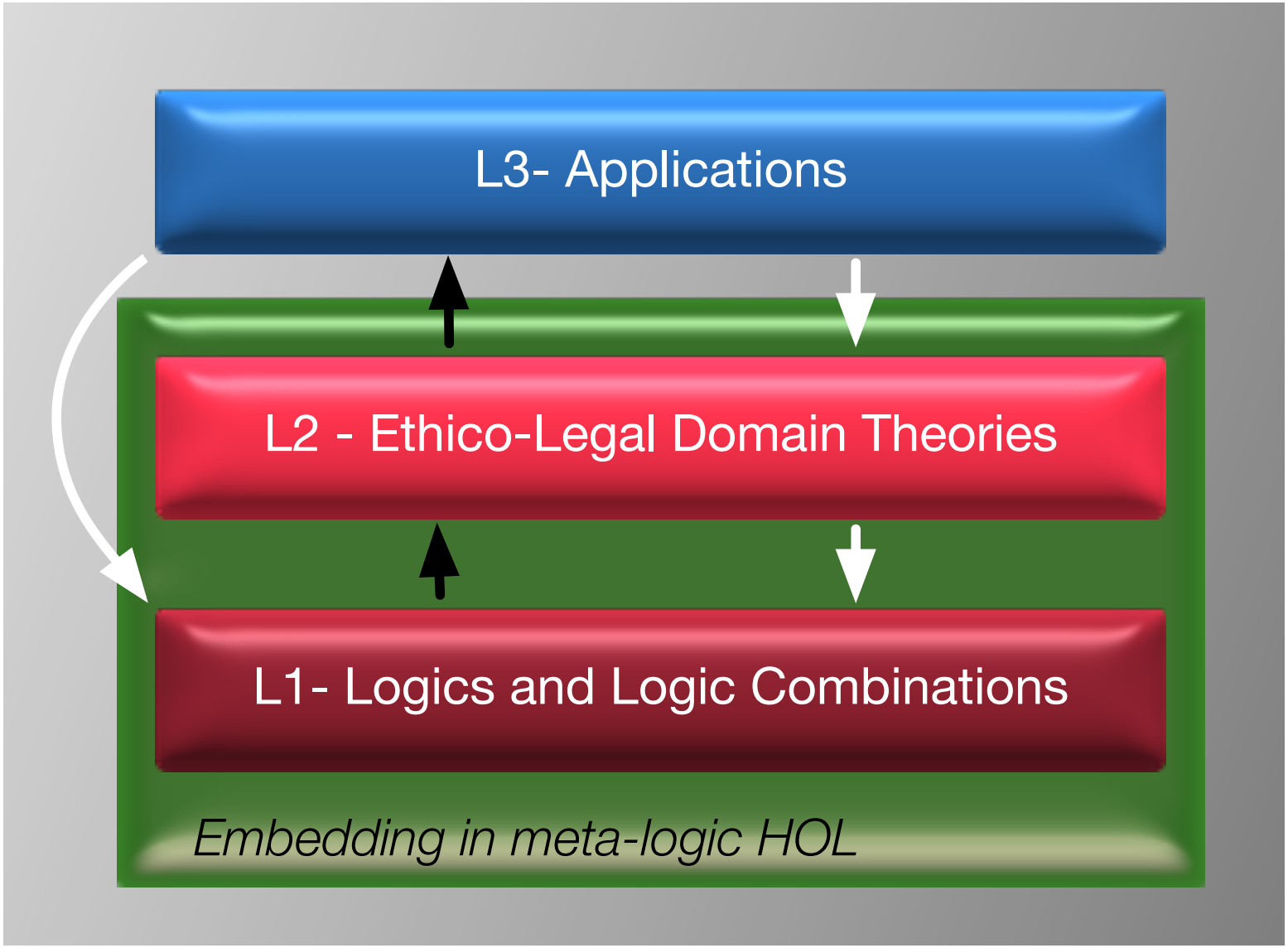


Normative Reasoning Experimentation Platform



LogiKEy Logic and Knowledge Engineering Methodology

[BenzmüllerParentTorre, arXiv:1903.10187, 2019]



Library of Deontic Logics

Base Logics

- Standard Deontic Logic
- Daydic Deontic Logic [Carmo& Jones, 2013]
- Aqvist System E [Aqvist, 2002]
- Input/Output Logic [Makinson&VanDerTorre, 2002]

Extensions

- first-order
- higher-order
- possibilist quantifiers (constant domain)
- actualist quantifiers (varying domain)

Combinations with other Logics

- Kaplan's Logic of Demonstratives (LD)
- fusion
- product

Library of Deontic Logics

Base Logics

- Standard Deontic Logic
- Daydic Deontic Logic [Carmo& Jones, 2013]
- Aqvist System E [Aqvist, 2002]
- Input/Output Logic [Makinson&VanDerTorre, 2002]

Extensions

- first-order
- higher-order
- possibilist quantifiers (constant domain)
- actualist quantifiers (varying domain)

Combinations with other Logics

- Kaplan's Logic of Demonstratives (LD)
- fusion
- product

Library of Deontic Logics

Base Logics

- Standard Deontic Logic
- Daydic Deontic Logic [Carmo& Jones, 2013]
- Aqvist System E [Aqvist, 2002]
- Input/Output Logic [Makinson&VanDerTorre, 2002]

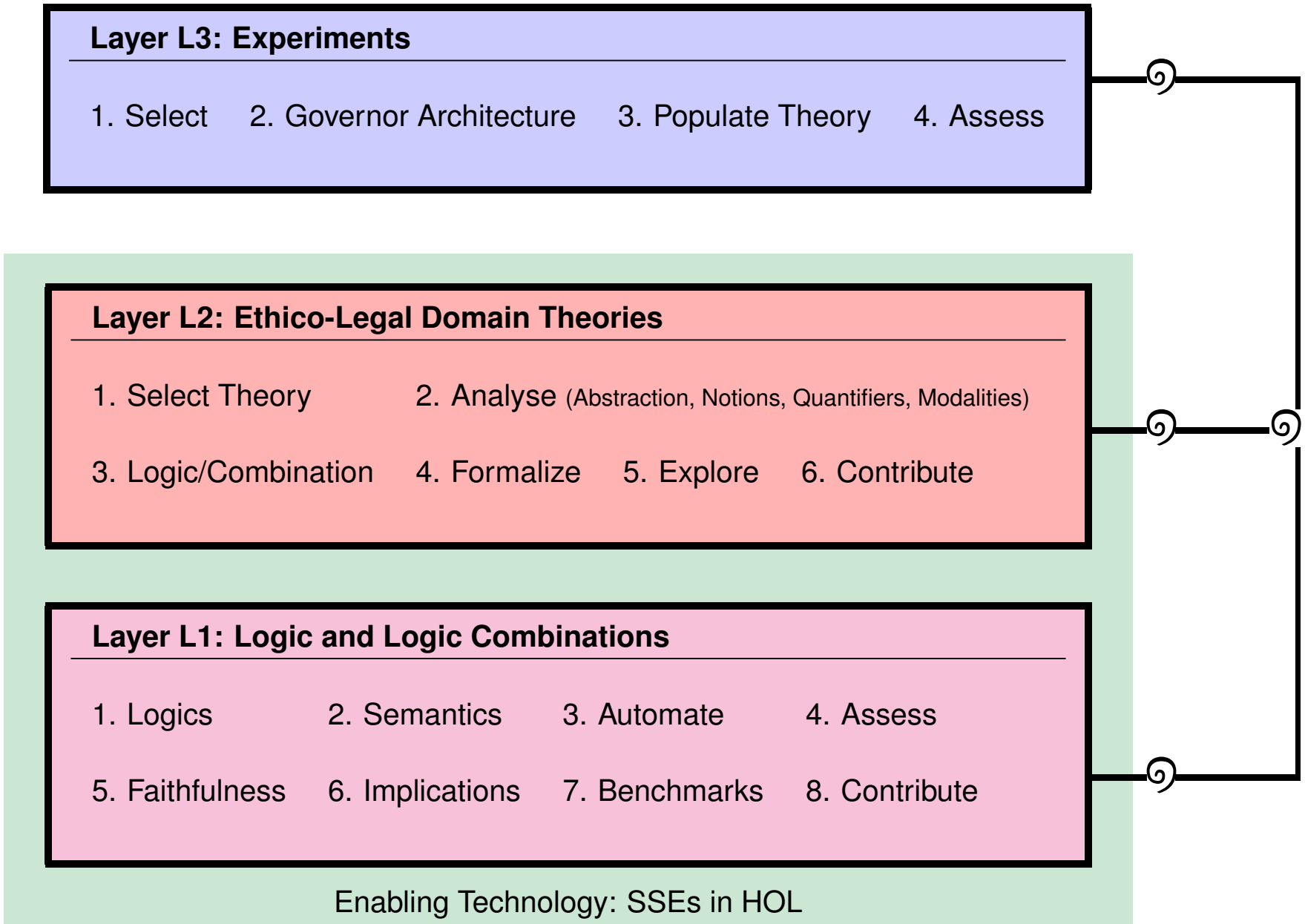
Extensions

- first-order
- higher-order
- possibilist quantifiers (constant domain)
- actualist quantifiers (varying domain)

Combinations with other Logics

- Kaplan's Logic of Demonstratives (LD)
- fusion
- product

LogiKEy Logic and Knowledge Engineering Methodology



What's Next?

What's Next?

Proposal: **Laboratory on Machine Ethics**

- **Experiments in Normative Reasoning and Machine-Ethics**
  - Cloud-based Plattform for Universal Normative Reasoning
  - Ethico-legal Governance of Autonomous Systems
- **Transdisciplinary:** KR&R, Formal Ethics/Philosophy, Law, Machine Learning
- **International Team** & **Industry Particpipation**

Let’s start now working with Isabelle

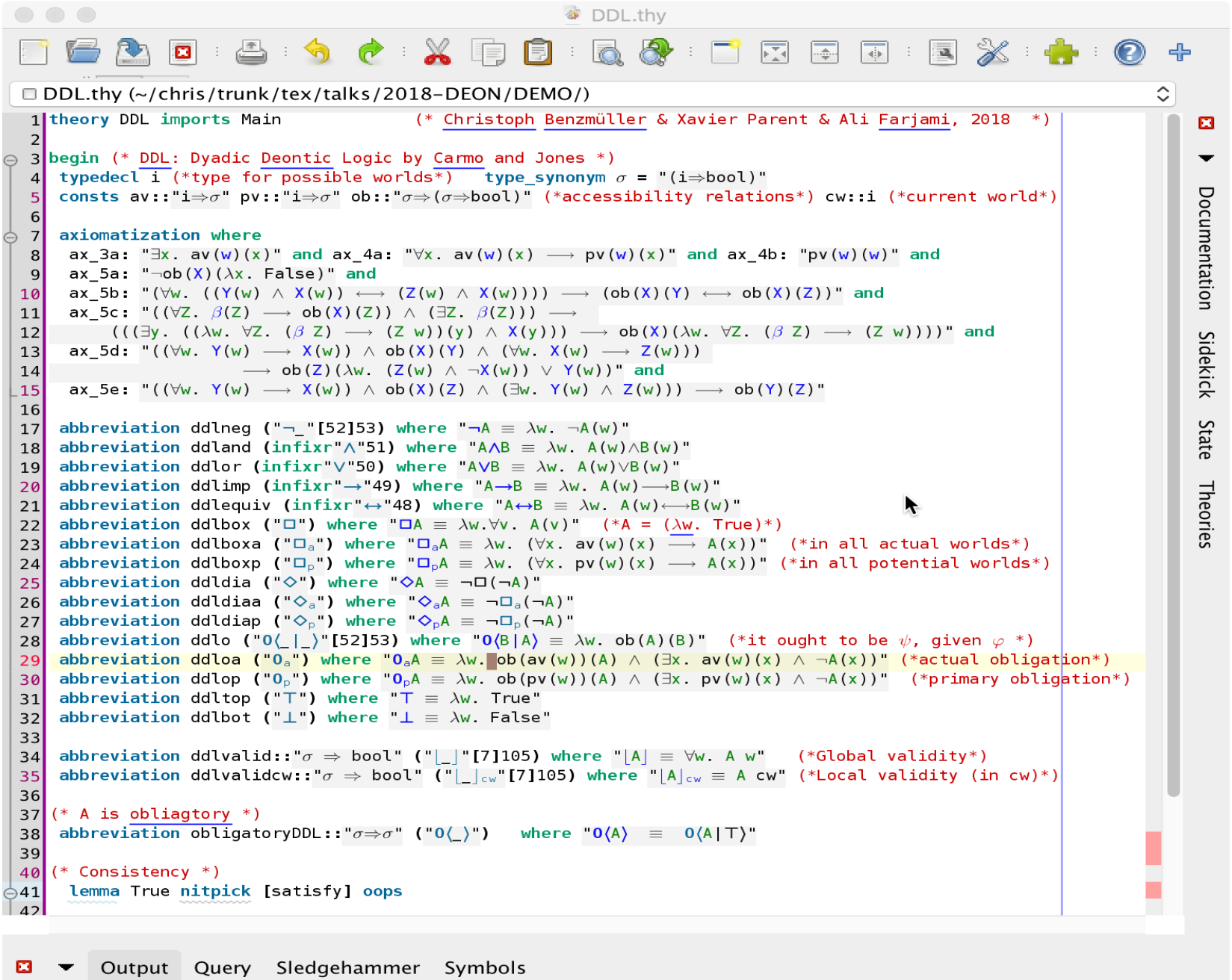
Demo: SDL in Isabelle/HOL

[Logica Universalis, 2013]

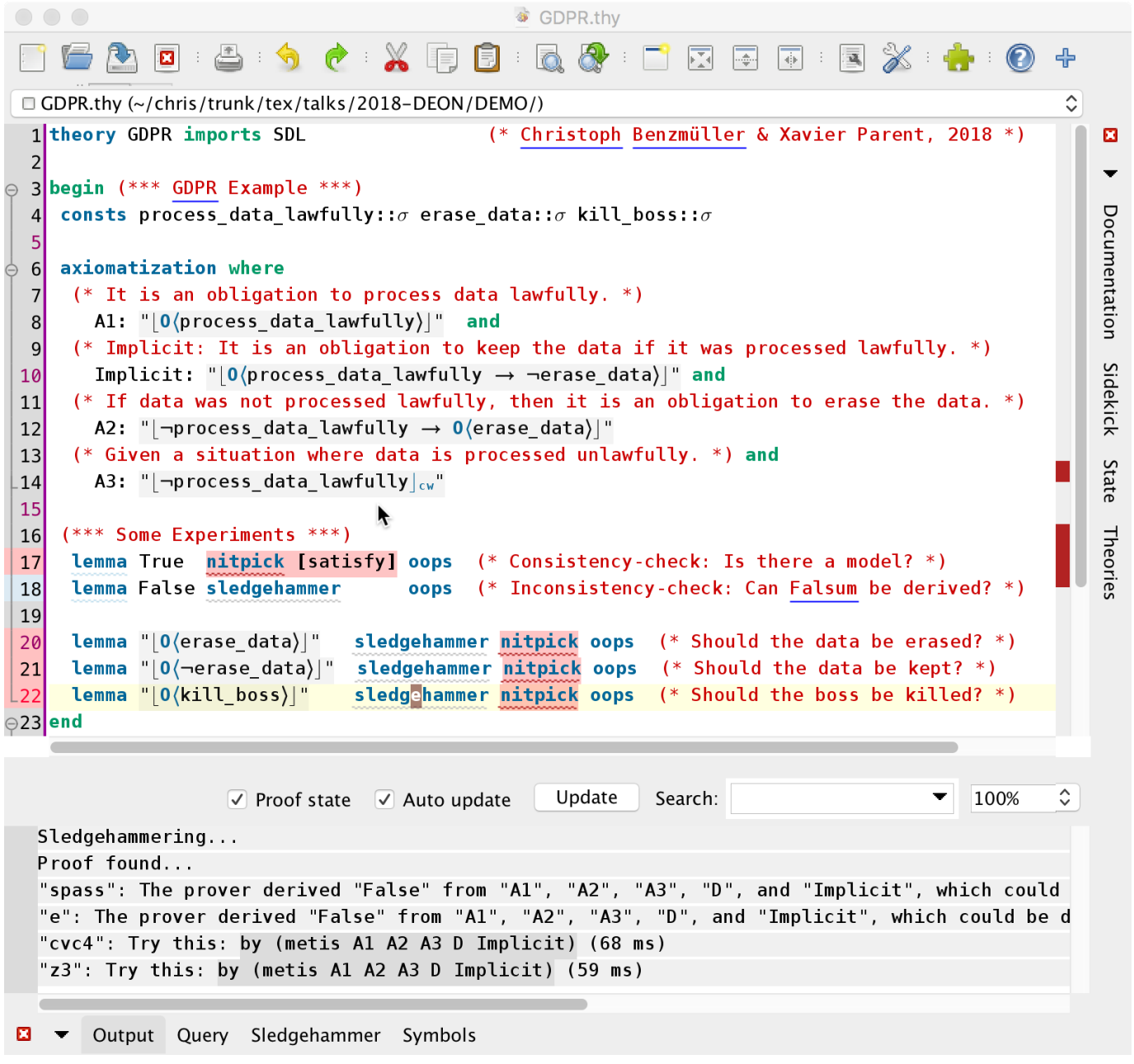
Isabelle2018/HOL - SDL.thy

```
theory SDL imports Main  (* SDL: Standard Deontic Logic. C. Benzmüller & X. Parent, 2019 *)
begin
 typedecl i (*Type for possible worlds.*)  type_synonym σ = "(i⇒bool)"
 type_synonym γ = "σ⇒σ" type_synonym ϱ = "σ⇒σ⇒σ"
 consts R::"i⇒i⇒bool" (infixr"R"70) (*Accessibility relation.*)  aw::i (*Actual world.*)

 abbreviation SDLtop::σ ("⊤") where "⊤ ≡ λw. True"
 abbreviation SDLbot::σ ("⊥") where "⊥ ≡ λw. False"
 abbreviation SDLnot::γ ("¬_"[52]53) where "¬φ ≡ λw. ¬φ(w)"
 abbreviation SDLand::ϱ (infixr"∧"51) where "φ∧ψ ≡ λw. φ(w) ∧ ψ(w)"
 abbreviation SDLor::ϱ (infixr"∨"50) where "φ∨ψ ≡ λw. φ(w) ∨ ψ(w)"
 abbreviation SDLimp::ϱ (infixr"→"49) where "φ→ψ ≡ λw. φ(w) ⟶ ψ(w)"
 abbreviation SDLequ::ϱ (infixr"↔"48) where "φ↔ψ ≡ λw. φ(w) ⟷ ψ(w)"

(*Possibilist Quantification.*)
 abbreviation mforall ("∀") where "∀Φ ≡ λw.∀x. (Φ x w)"
 abbreviation mforallB (binder"∀"[8]9) where "∀x. φ(x) ≡ ∀φ"
 abbreviation mexists ("∃") where "∃Φ ≡ λw.∃x. (Φ x w)"
 abbreviation mexistsB (binder"∃"[8]9) where "∃x. φ(x) ≡ ∃φ"

(*Obligation*)
 abbreviation SDLobligatory::γ ("OB") where "OB φ ≡ λw. ∀v.  w R v ⟶ φ(v)"
 abbreviation SDLpermissible::γ ("PE") where "PE φ ≡ ¬(OB(¬φ))"

(*Validity*)
 abbreviation SDLvalid::"σ⇒bool" ("⌊_⌋"[7]105)      (*Global validity.*)
   where "⌊A⌋ ≡ ∀w. A w"
 abbreviation SDLvalidcw::"σ⇒bool" ("⌊_⌋ι"[7]105)   (*Validity in actual world.*)
   where "⌊A⌋ι ≡ A aw"

(*SDL axiom D*)
 axiomatization where serial_R: "∀x. ∃y. x R y" (*Accessibility relation R is serial*)

 lemma D:  "⌊∀φ. ¬ ((OB φ) ∧ (OB (¬ φ)))⌋" using serial_R by auto (*Axiom D*)
 lemma D': "⌊∀φ.(OB φ) ⟶ (PE φ)⌋"          using serial_R by auto (*Axiom D'*)

(*Notation*)
 abbreviation SDLobl::γ ("O<_>") where "O<A> ≡  OB A"  (*New syntax: A is obligatory.*)

(*Consistency confirmed by model finder Nitpick.*)
 lemma True nitpick[satisfy,user_axioms,expect=genuine,show_all,format=2] oops

(*Barcan Formulas*)
 lemma Barcan:         "⌊(∀d. O<φ(d)>) ⟶ (O<∀d. φ(d)>)⌋" by simp
 lemma ConverseBarcan: "⌊(O<∀d. φ(d)>) ⟶ (∀d. O<φ(d)>)⌋" by simp
end
```

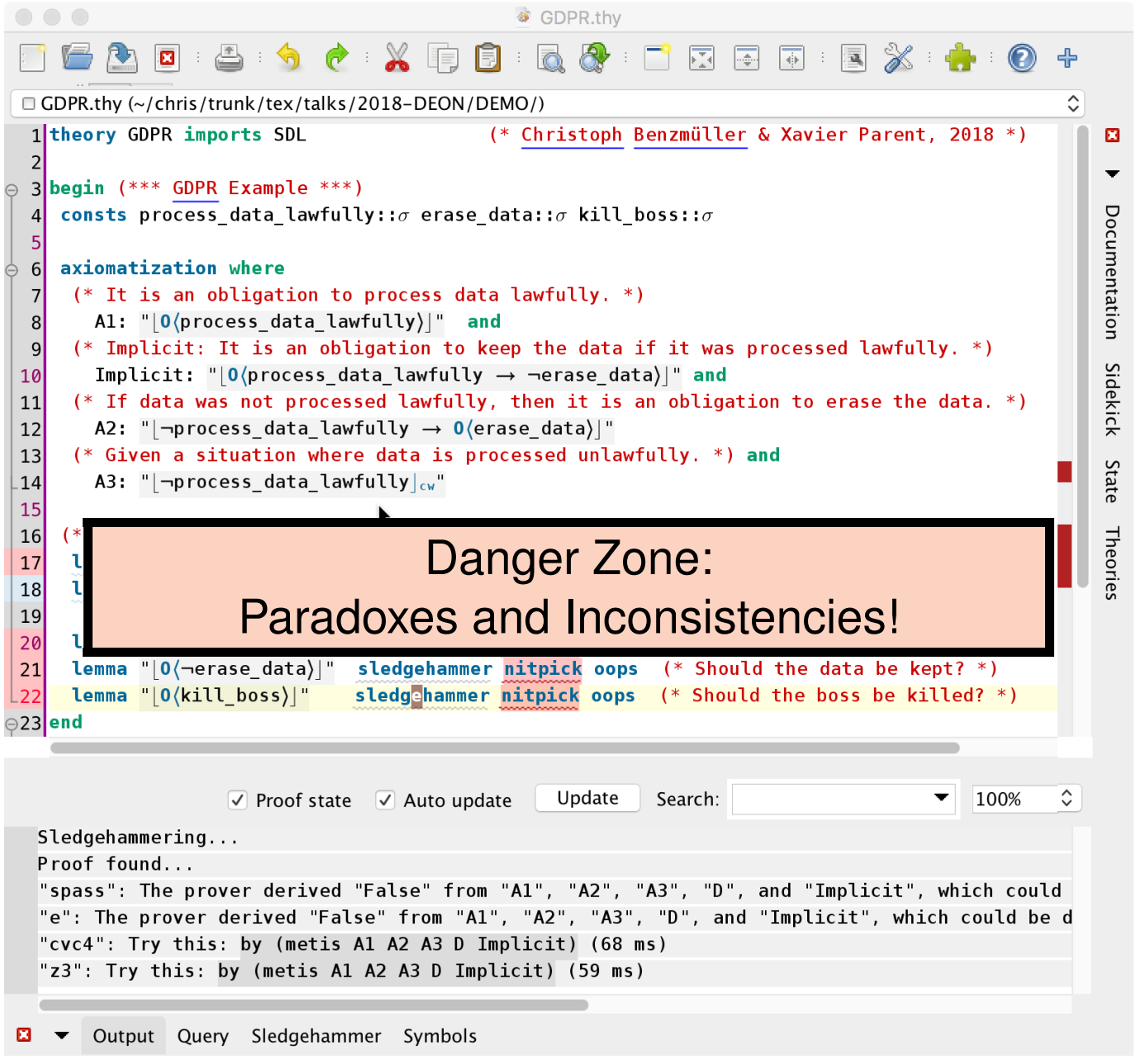
Demo: DDL in Isabelle/HOL



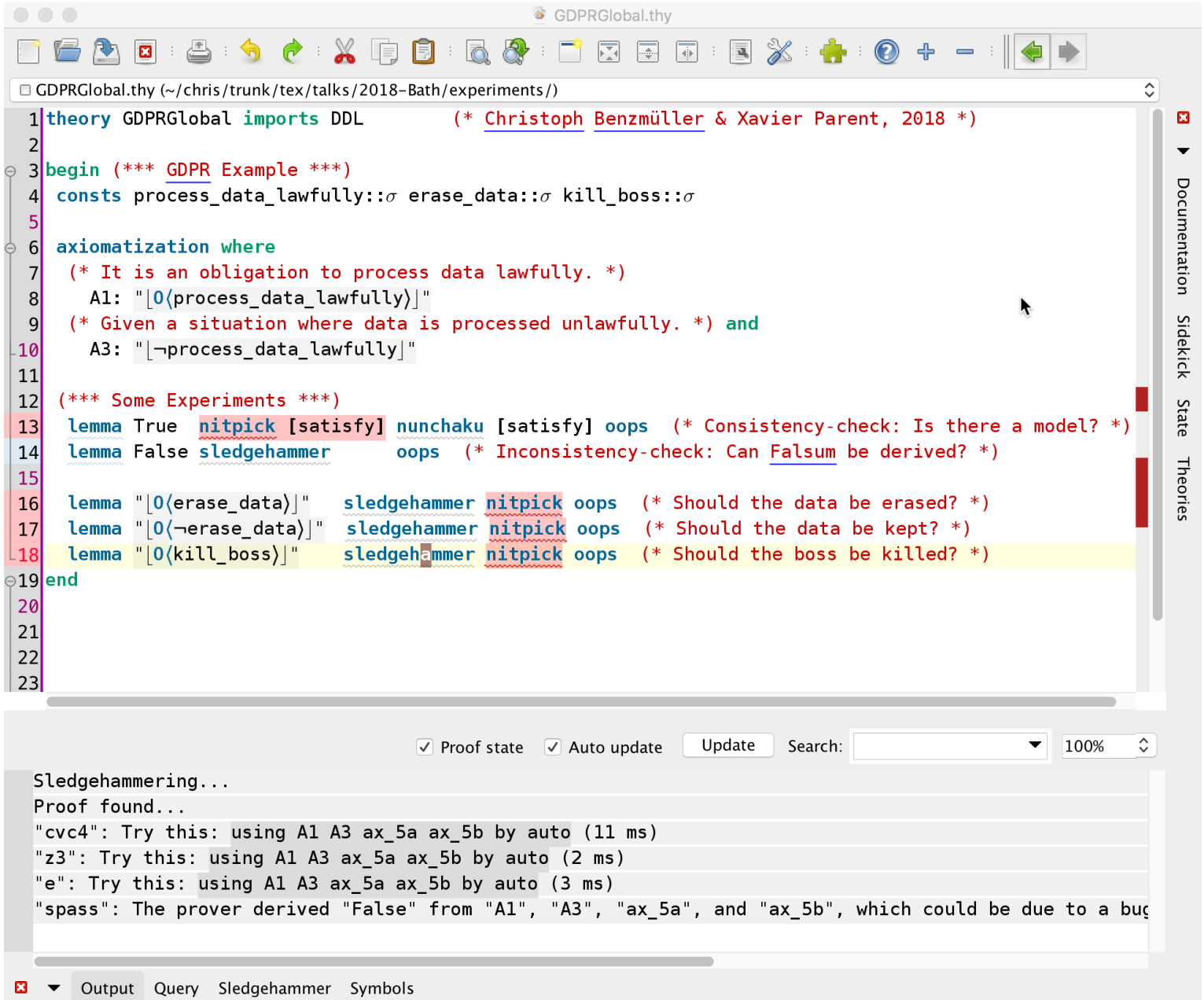
Demo: Normative Reasoning Experimentation Platform



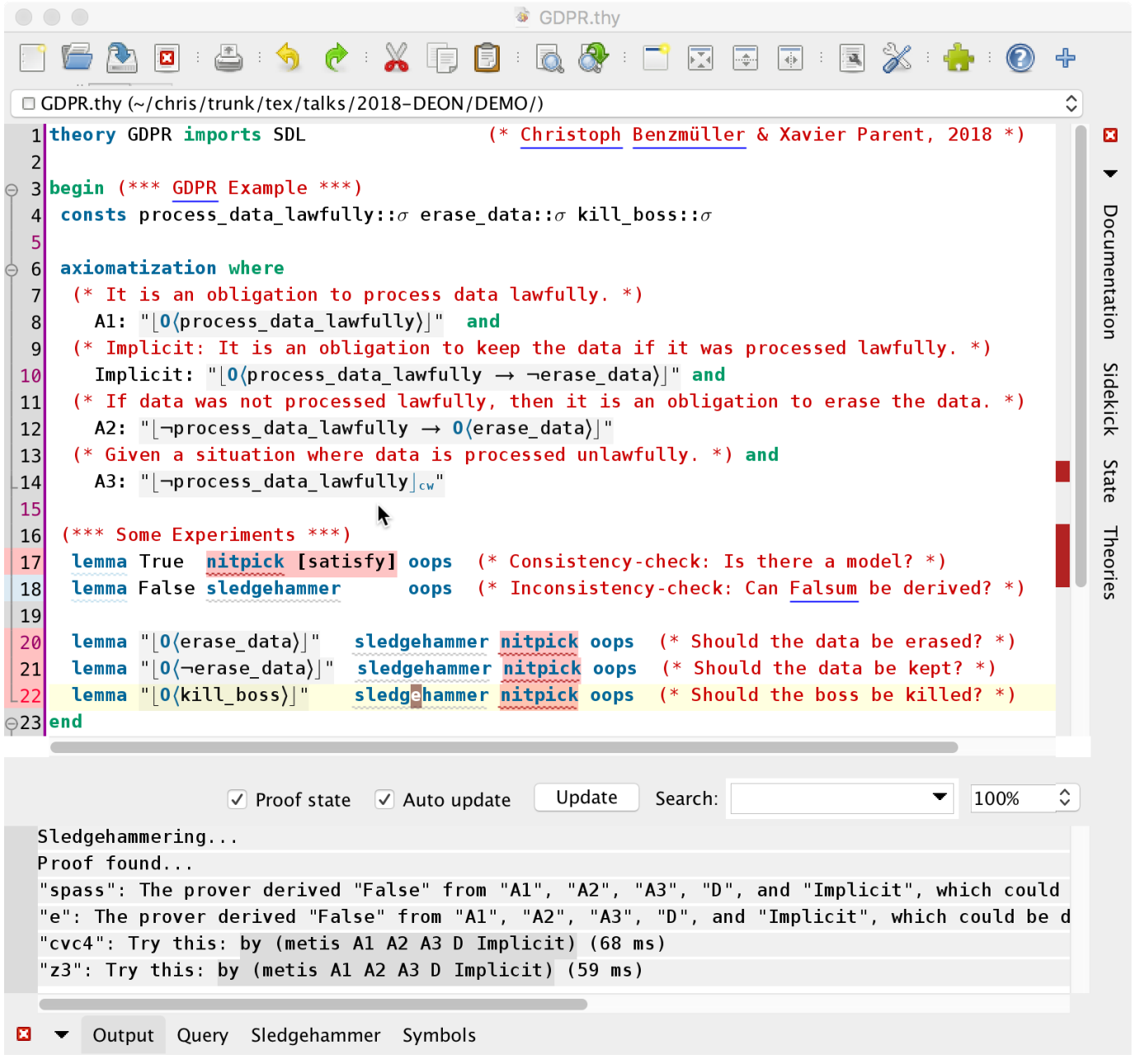
Demo: Normative Reasoning Experimentation Platform



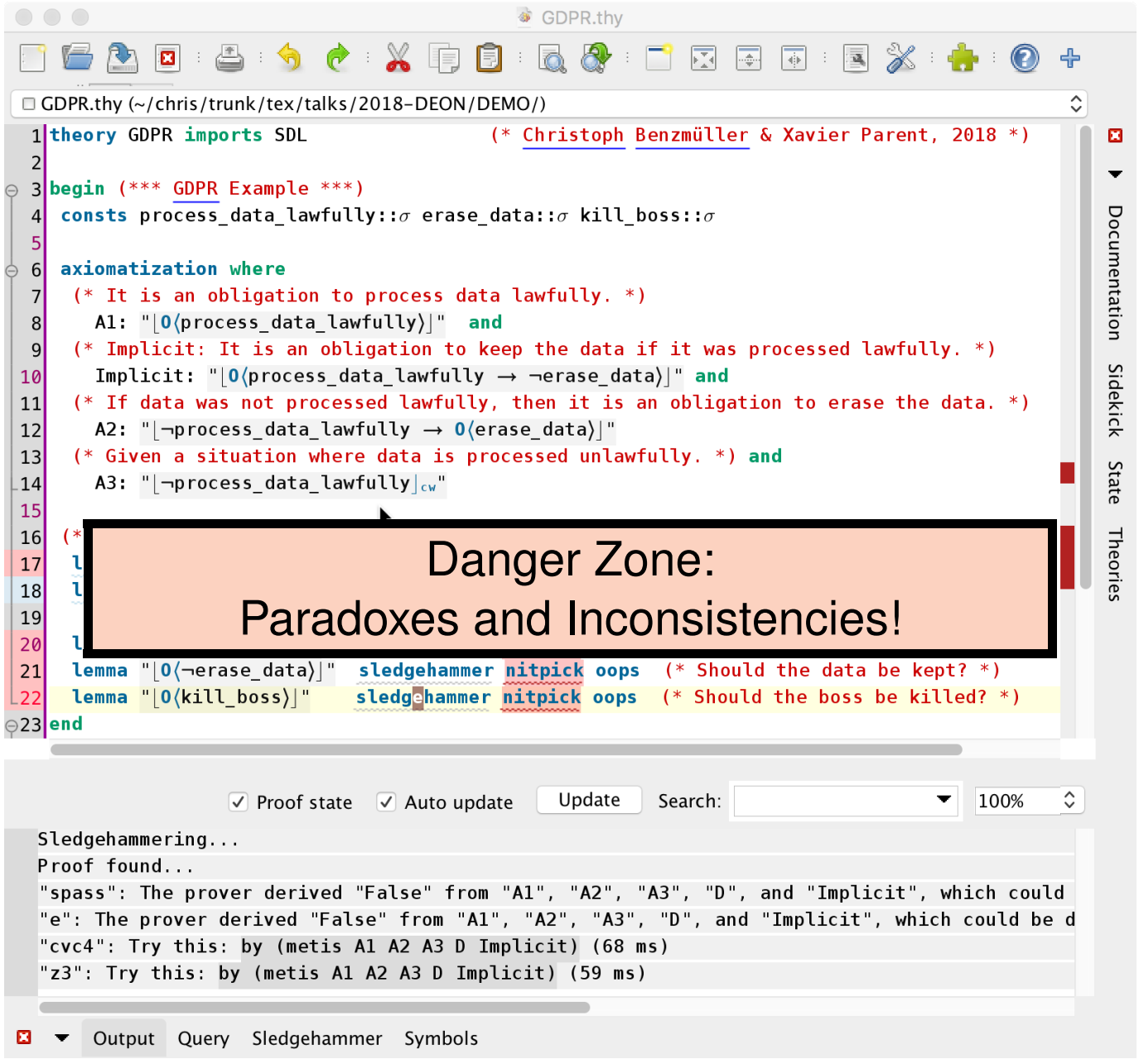
Demo: Global vs. Local Consequence Relation



Demo: Normative Reasoning Experimentation Platform



Demo: Normative Reasoning Experimentation Platform



Gewirth's PGC in Isabelle/HOL

[Fuenmayor&B., PRICAI & FroCoS, 2019]

> "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 intellig. (AGI)
- **Idea (in a nutshell):**
  - emendation of the Golden Rule (treat others as you would wish to be treated)
  - adopting an agent perspective, the PGC expresses, and deductively justifies, a related upper moral principle, according to which any intelligent agent, by virtue of its self-understanding as an agent, is rationally committed to asserting that:
    (i) it has rights to freedom and well-being, and
    (ii) all other agents have those same rights.
- **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. UCP 1991.
  - A. Kornai. Bounding the impact of AGI. J. Experimental & Theoretical AI, 2014.

Gewirth's PGC in Isabelle/HOL

[Fuenmayor&B., PRICAI & FroCoS, 2019]

> "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 intellig. (AGI)
- **Idea (in a nutshell):**
  - emendation of the Golden Rule (treat others as you would wish to be treated)
  - adopting an agent perspective, the PGC expresses, and deductively justifies, a related upper moral principle, according to which any intelligent agent, by virtue of its self-understanding as an agent, is rationally committed to asserting that:
    (i) it has rights to freedom and well-being, and
    (ii) all other agents have those same rights.
- **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. UCP 1991.
  - A. Kornai. Bounding the impact of AGI. J. Experimental & Theoretical AI, 2014.

Gewirth's PGC in Isabelle/HOL

[Fuenmayor&B., PRICAI & FroCoS, 2019]

> "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 intellig. (AGI)
- **Idea (in a nutshell):**
  - emendation of the Golden Rule (treat others as you would wish to be treated)
  - adopting an agent perspective, the PGC expresses, and deductively justifies, a related upper moral principle, according to which any intelligent agent, by virtue of its self-understanding as an agent, is rationally committed to asserting that:
    (i) it has rights to freedom and well-being, and
    (ii) all other agents have those same rights.
- **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. UCP 1991.
  - A. Kornai. Bounding the impact of AGI. J. Experimental & Theoretical AI, 2014.

Gewirth's PGC in Isabelle/HOL

[Fuenmayor&B., PRICAI & FroCoS, 2019]

> "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 intellig. (AGI)
- **Idea (in a nutshell):**
  - emendation of the Golden Rule (treat others as you would wish to be treated)
  - adopting an agent perspective, the PGC expresses, and deductively justifies, a related upper moral principle, according to which any intelligent agent, by virtue of its self-understanding as an agent, is rationally committed to asserting that:
    (i) it has rights to freedom and well-being, and
    (ii) all other agents have those same rights.
- **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. UCP 1991.
  - A. Kornai. Bounding the impact of AGI. J. Experimental & Theoretical AI, 2014.

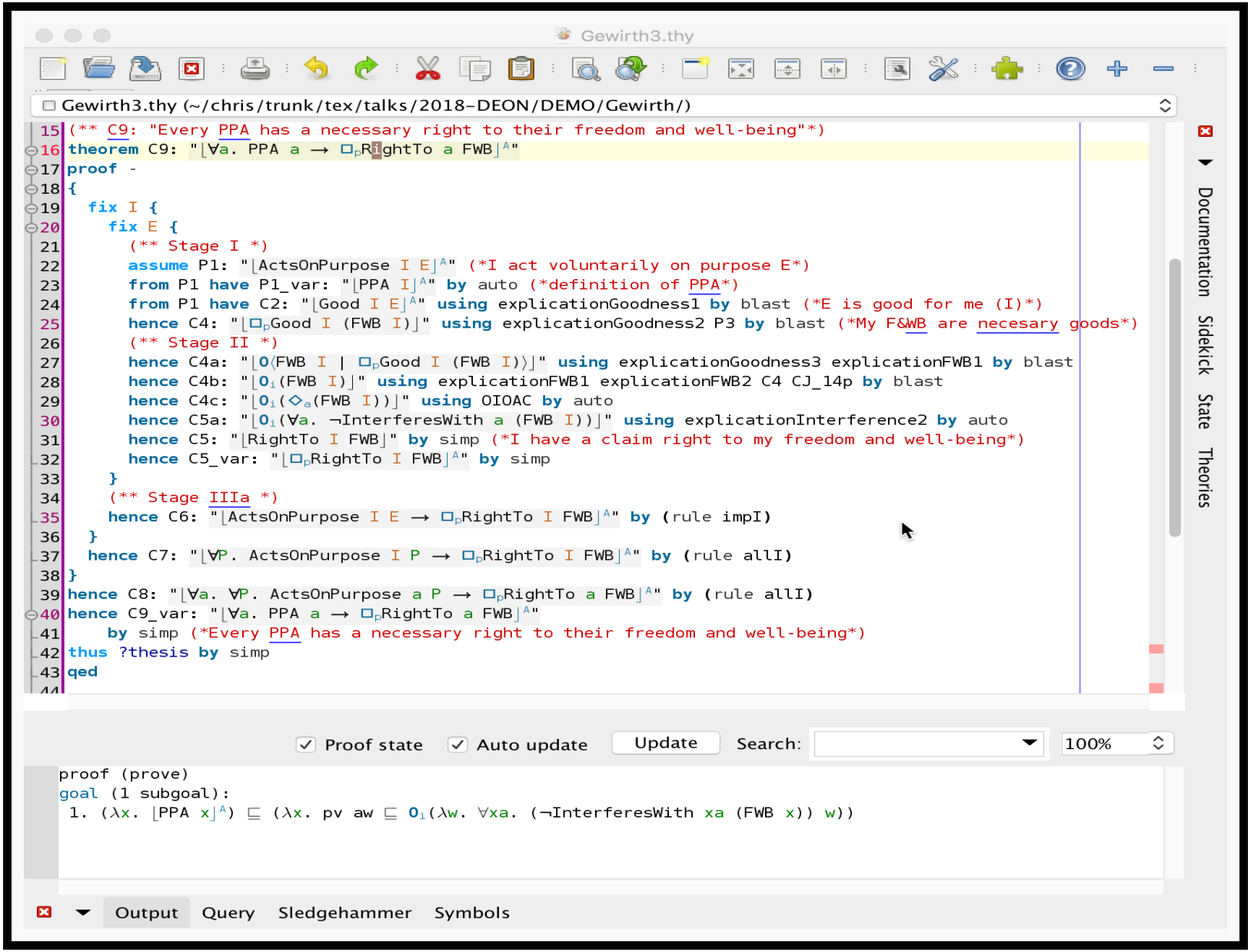
Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

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**

Gewirth's PGC in Isabelle/HOL

[Fuenmayor&B., PRICAI & FroCoS, 2019]



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

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL

[Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents)
  an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$
  a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents)
  an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$
  a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \ \wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL [Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Gewirth's PGC in Isabelle/HOL

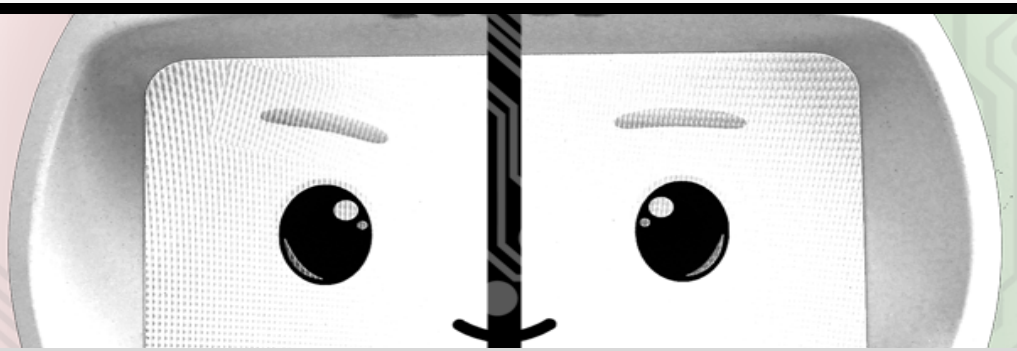
[Fuenmayor&B., PRICAI & FroCoS, 2019]

- challenging: **alethic** & **deontic modalities**, **quantification** and **indexicals**
- uninterpreted predicate/relation symbols: `FWB`, `Good` (1-ary); `ActsOnPurpose`, `InterferesWith` (both 2-ary); `NeedsForPurpose` (3-ary)
- $\texttt{PPA}\,a := \exists E.\,\texttt{ActsOnPurpose}\,a\,E$ (prospective purposive agents) an additional axiom postulates that being a `PPA` is identity-constitutive for any individual: $\lfloor \forall a.\texttt{PPA}a \rightarrow \Box^D(\texttt{PPA}a) \rfloor^D$.
- $\lfloor \varphi \rfloor^D$ is indexical validity of φ: true in all contexts (following Kaplan's LD)
- $\texttt{RightTo}\,a\,\varphi := \mathbf{O}_i(\forall b.\neg\texttt{InterferesWith}\,b\,(\varphi\,a))$ a has (claim) right to property φ iff it is obligatory that every b does not interfere with state of affairs $(\varphi\,a)$
- $\mathbf{O}_i$ is *ideal* obligation operator from Carmo & Jones' DDL ($\mathbf{O}_a$ also fine)
- further axioms interrelate the concepts of goodness and agency
- axiom $\lfloor \forall P.\forall a.\,\texttt{NeedsForPurpose}\,a\,\texttt{FWB}\,P \rfloor^D$ expresses that `FWB` is always required in order to act on any purpose
- `FWB` is postulated a contingent property: $\lfloor \forall a.\Diamond_p\,\texttt{FWB}\,a \;\wedge \Diamond_p \neg\texttt{FWB}\,a \rfloor^D$
- Note that both **first-order** and **higher-order quantifiers** are required

Ethics

Argued for explicit ethical reasoning competencies in IASs

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



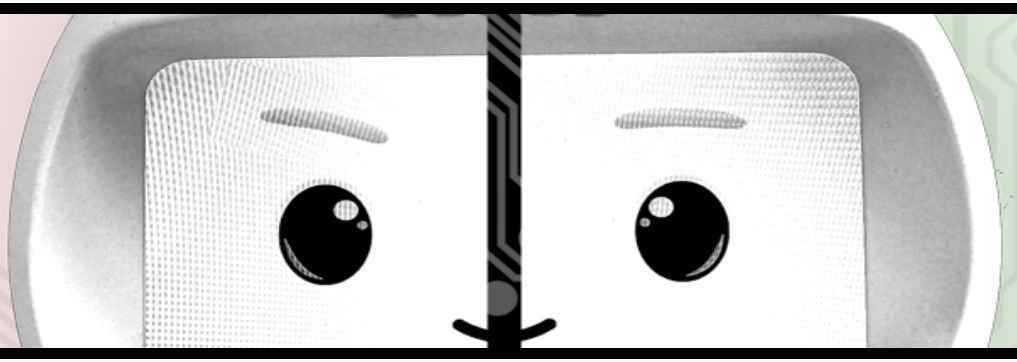
Ongoing and further work

- workbench of (expressive) deontic logics and logic combinations
- formalisation and mechanisation of ethico-legal theories
- experimentation and deployment in concrete applications

Ethics

Argued for explicit ethical reasoning competencies in IASs

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



Ongoing and further work

- workbench of (expressive) deontic logics and logic combinations
- formalisation and mechanisation of ethico-legal theories
- experimentation and deployment in concrete applications