

Ethisch-rechtliche Kontrolle von KI Systemen

Christoph Benz Müller

Freie Universität Berlin | Universität Luxembourg



Bitkom AI Research Network, bitkom.org, 21. Februar, 2020



- ▶ Wissen unsere heutigen (flachen) KI-Systeme was sie tun?
- ▶ Wissen wir was wir tun, wenn wir solchen KI-Systemen zunehmend kritische Entscheidungen übertragen?
- ▶ Ist die normative Richtungslosigkeit und Unberechenbarkeit Kerncharakter zukünftiger intelligenter autonomer Systeme?

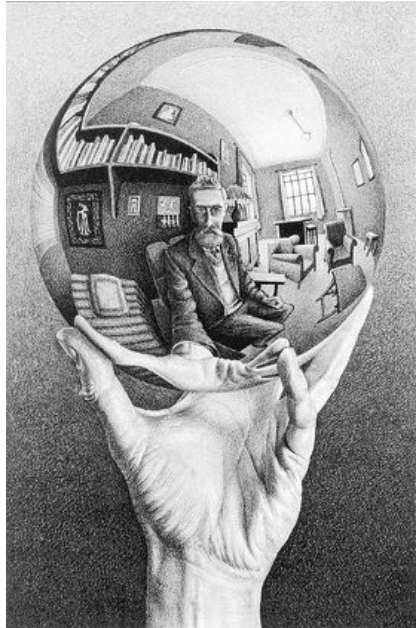


- ▶ Wissen unsere heutigen (flachen) KI-Systeme was sie tun?
- ▶ Wissen wir was wir tun, wenn wir solchen KI-Systemen zunehmend kritische Entscheidungen übertragen?
- ▶ Ist die normative Richtungslosigkeit und Unberechenbarkeit Kerncharakter zukünftiger intelligenter autonomer Systeme?



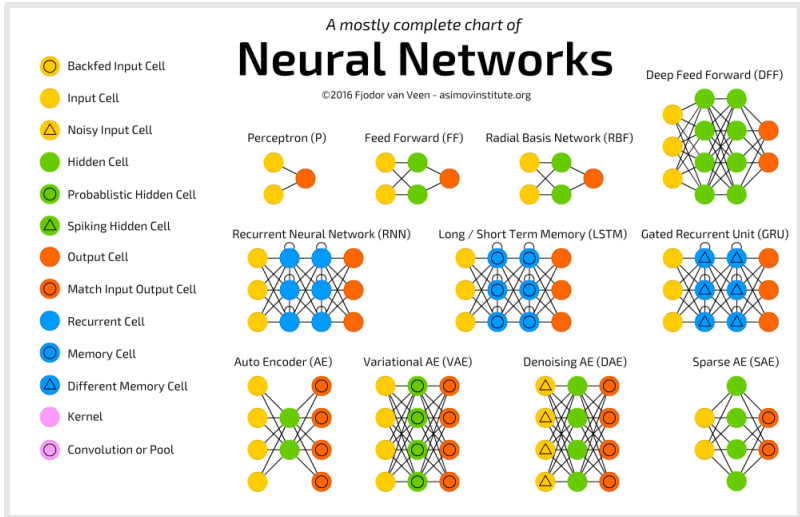
- ▶ Wissen unsere heutigen (flachen) KI-Systeme was sie tun?
- ▶ Wissen wir was wir tun, wenn wir solchen KI-Systemen zunehmend kritische Entscheidungen übertragen?
- ▶ Ist die normative Richtungslosigkeit und Unberechenbarkeit Kerncharakter zukünftiger intelligenter autonomer Systeme?

- ▶ Introspektion & Selbstreflektion?
- ▶ Wie ist es also um die “Mündigkeit” heutiger KI-Systeme bestellt?
- ▶ Bürgerrechte verleihen?



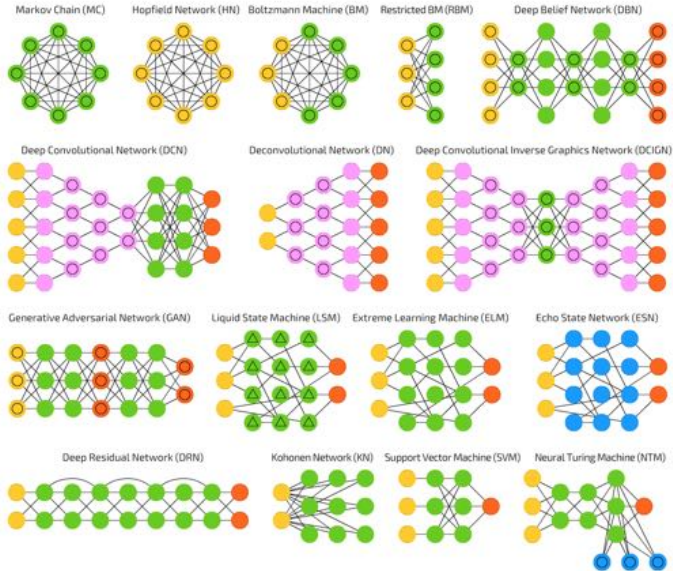
Tiefe Neuronale Netzwerke und Datengetriebene KI

Quelle: Fjodor van Veen, asimovinstitute.org, 2016

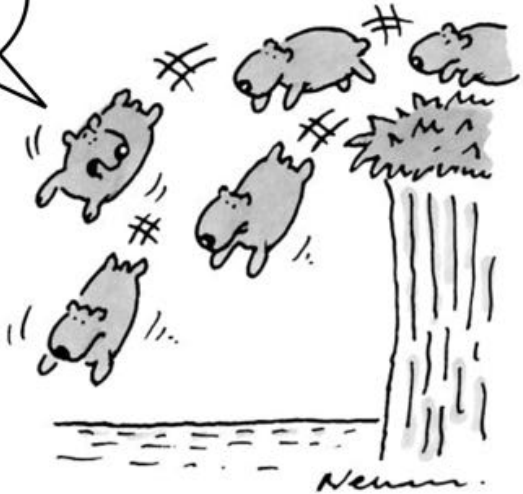


Tiefe Neuronale Netzwerke und Datengetriebene KI

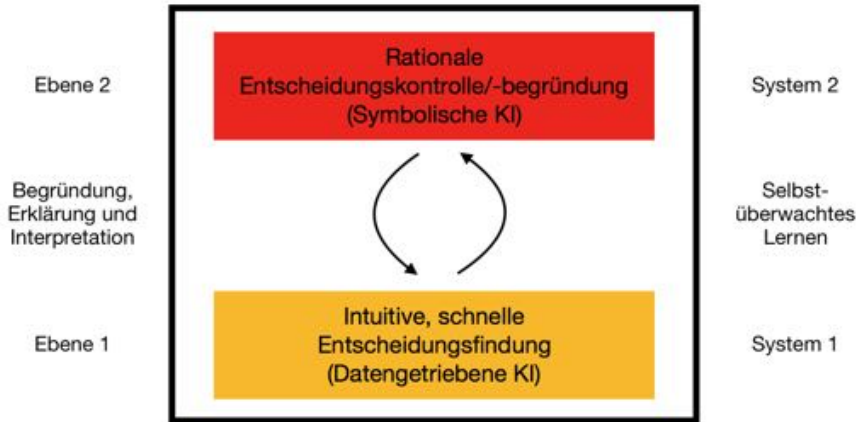
Quelle: Fjodor van Veen, asimovinstitute.org, 2016



Who claimed that
"Data Driven AI"
is the answer to all
challenges?



Intuitives vs. Rationales Schließen



Vgl. auch Kahnemann, "Thinking, fast and slow", 2013

Intuitives vs. Rationales Schließen

Quelle: Jonathan Haidt, The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment, Psychological Review, 2001

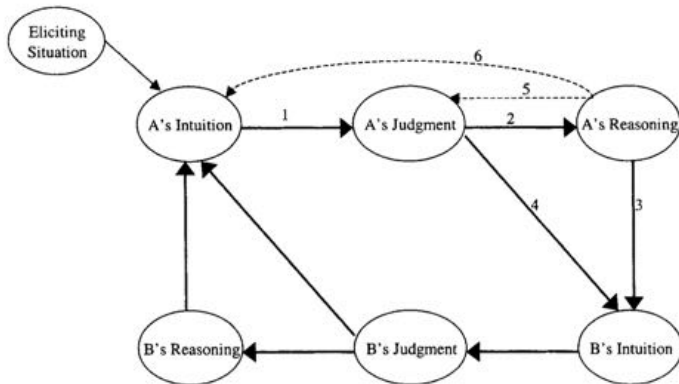


Figure 2. The social intuitionist model of moral judgment. The numbered links, drawn for Person A only, are (1) the intuitive judgment link, (2) the post hoc reasoning link, (3) the reasoned persuasion link, and (4) the social persuasion link. Two additional links are hypothesized to occur less frequently: (5) the reasoned judgment link and (6) the private reflection link.

Eigene Arbeitsdefinition von KI

Benzmüller (März 2019)

Def.: Künstliche Intelligenz

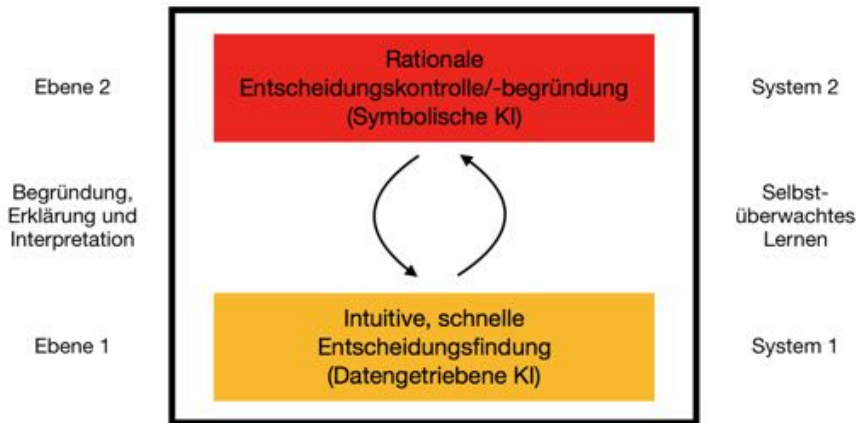
Eine Wissenschaft komputationaler Technologien, welche entwickelt werden um *intelligentes* Verhalten in Maschinen zu erreichen und zu erklären.

Def.: Intelligenz

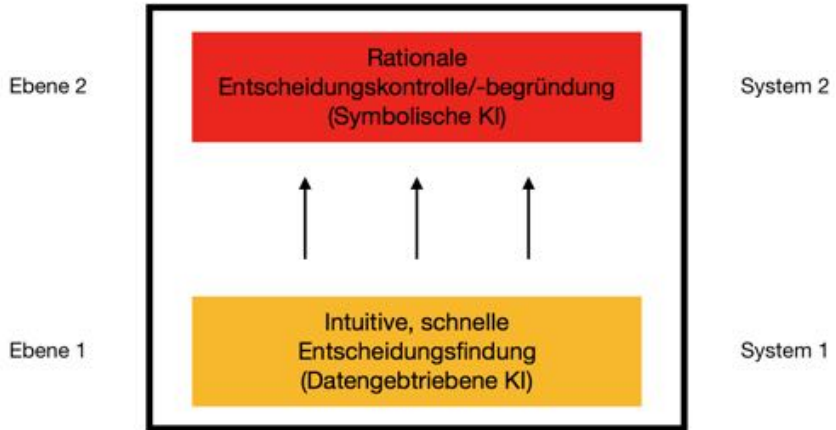
Eine Kollektion mentaler Fähigkeiten die eine Entität dazu befähigen

1. schwierige **Probleme** zu **lösen** (oder zu lernen wie man diese löst),
2. das **Unbekannte** zu **meistern**: erfolgreich in bekannten, unbekannten und dynamischen Umgebungen zu agieren (Wahrnehmung, Planen, etc.),
3. **abstrakt und rational** zu **schließen**, unter Vermeidung von Widersprüchen,
4. über sich selbst zu **reflektieren** und das eigene Schließen auszurichten an übergeordneten Zielen und Normen, und
5. **sozial** zu **interagieren** mit anderen Entitäten und eigene Ziele und Normen anzupassen an die einer Gemeinschaft.

Langfristiges Ziel



Aktuelles Ziel



Unkritische vs. Kritische KI-Anwendungen



vs.



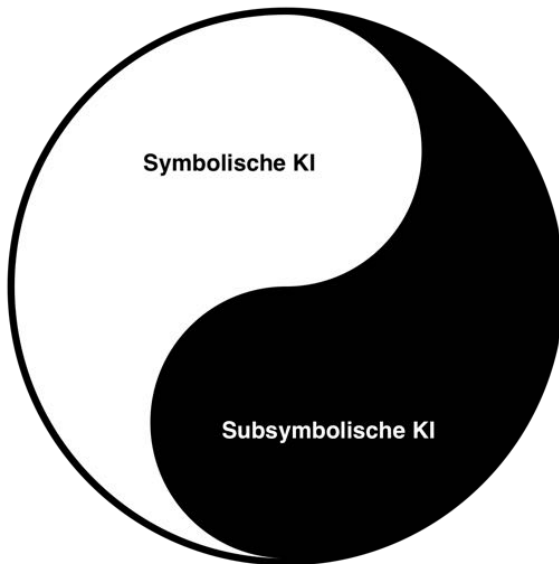
Links: Watson, the IBM-developed supercomputer, challenges Ken Jennings, left, and Brad Rutter to "Jeopardy!" in January 2011.

Rechts: The Defender, U.S. robotic platform that performs reconnaissance, surveillance, targeting, and threat neutralization tasks. Image: US Air Force



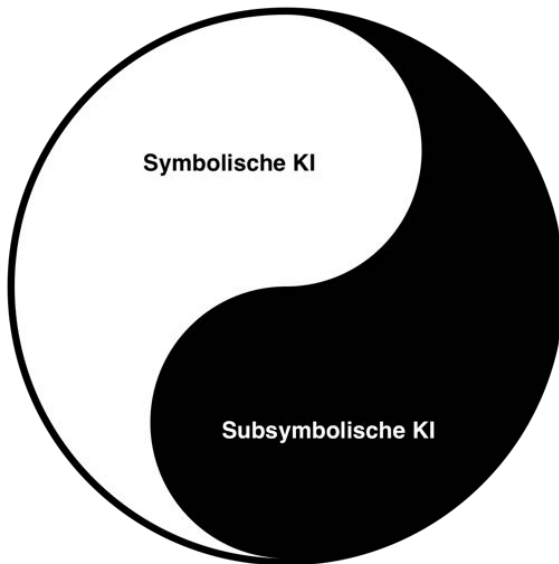
— Eigene Position, Motivation und Ziele —

Yin und Yang der KI— **Eigene Position**



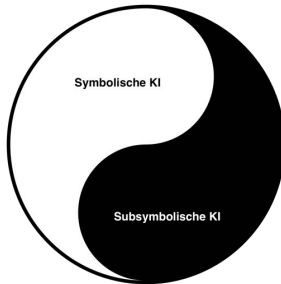
Forderung "Neuro-Symbolic AI" von Tim Cox (MIT-IBM Watson) oder Yoshua Bengio (Montreal) auf AAAI 2020.

Yin und Yang der KI— **Erfolge & Medienpräsenz**



Forderung "Neuro-Symbolic AI" von Tim Cox (MIT-IBM Watson) oder Yoshua Bengio (Montreal) auf AAAI 2020.

Yin und Yang der KI— **Erfolge & Medienpräsenz**

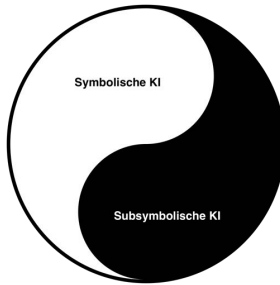


Erfolge

- ▶ Spiele:
Go und Schach
- ▶ endliche
Problemdomäne
- ▶ Weltmeisterniveau
erreicht: AlphaGo
und AlphaZero

(subsymbolische KI)

Yin und Yang der KI— **Erfolge & Medienpräsenz**



Erfolge

- ▶ Spiele:
Go und Schach
- ▶ endliche
Problemdomäne
- ▶ Weltmeisterniveau
erreicht: AlphaGo
und AlphaZero

(subsymbolische KI)

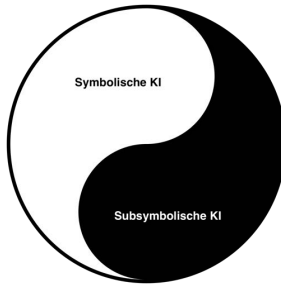


Yin und Yang der KI— **Erfolge & Medienpräsenz**

Erfolge

- ▶ Mathe:
Zahlentheorie
- ▶ unendliche
Problemdomäne
- ▶ offene Probleme
zuletzt gelöst durch
SAT-Solver

(symbolische KI)



Erfolge

- ▶ Spiele:
Go und Schach
- ▶ endliche
Problemdomäne
- ▶ Weltmeisterniveau
erreicht: AlphaGo
und AlphaZero

(subsymbolische KI)

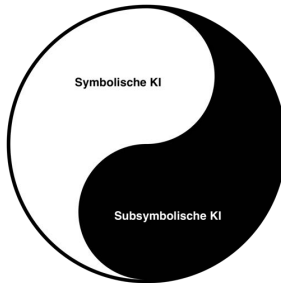


Yin und Yang der KI— **Erfolge & Medienpräsenz**

Erfolge

- ▶ Mathe:
Zahlentheorie
- ▶ unendliche
Problemdomäne
- ▶ offene Probleme
zuletzt gelöst durch
SAT-Solver

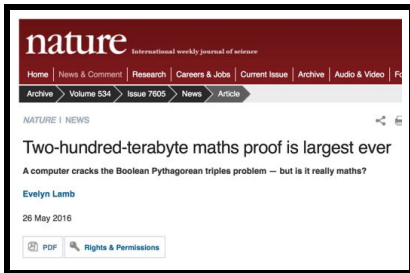
(symbolische KI)



Erfolge

- ▶ Spiele:
Go und Schach
- ▶ endliche
Problemdomäne
- ▶ Weltmeisterniveau
erreicht: AlphaGo
und AlphaZero

(subsymbolische KI)

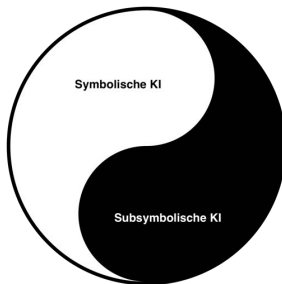


Yin und Yang der KI

Erfolge

- ▶ Mathe:
Zahlentheorie
- ▶ unendliche
Problemdomäne
- ▶ offene Probleme
zuletzt gelöst durch
SAT-Solver

(symbolische KI)



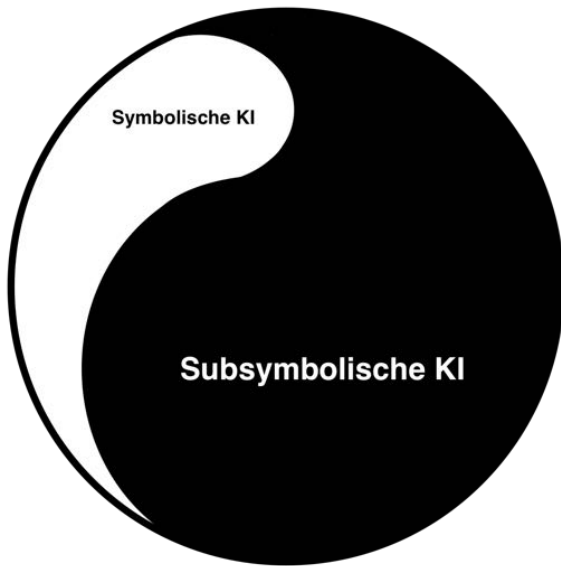
Erfolge

- ▶ Spiele:
Go und Schach
- ▶ endliche
Problemdomäne
- ▶ Weltmeisterniveau
erreicht: AlphaGo
und AlphaZero

(subsymbolische KI)



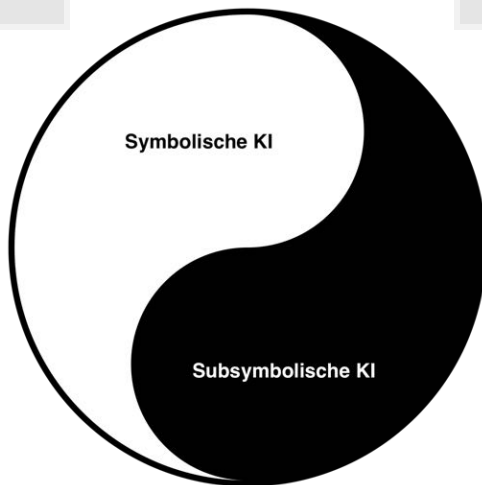
Yin und Yang der KI— **Ungesunder Hype?!**



Yin und Yang der KI — **The Next (really) Big Thing?!**

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

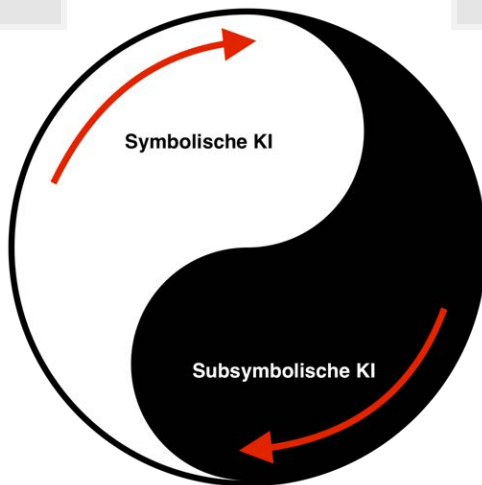
Korrelationen
Muster
Robustheit
Lernen
...



Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

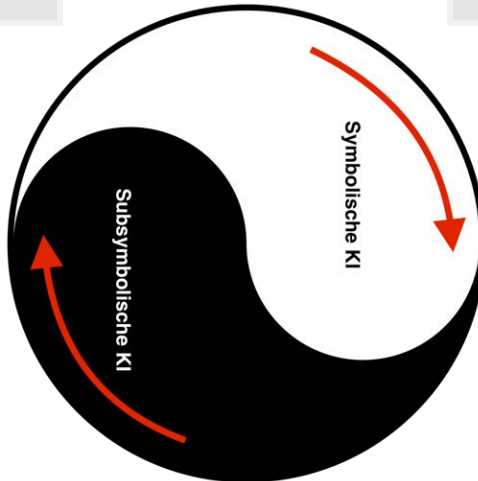
Korrelationen
Muster
Robustheit
Lernen
...



Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

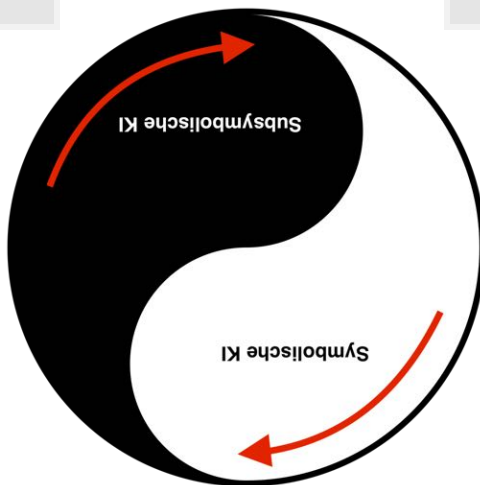
Korrelationen
Muster
Robustheit
Lernen
...



Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

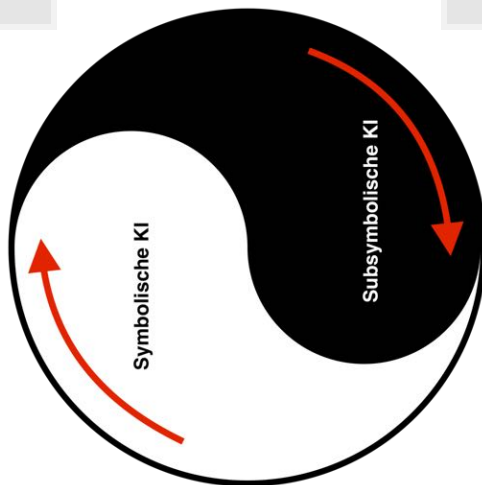
Korrelationen
Muster
Robustheit
Lernen
...



Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

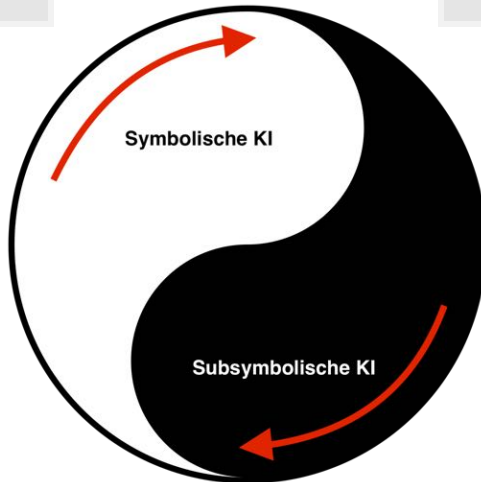
Korrelationen
Muster
Robustheit
Lernen
...



Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

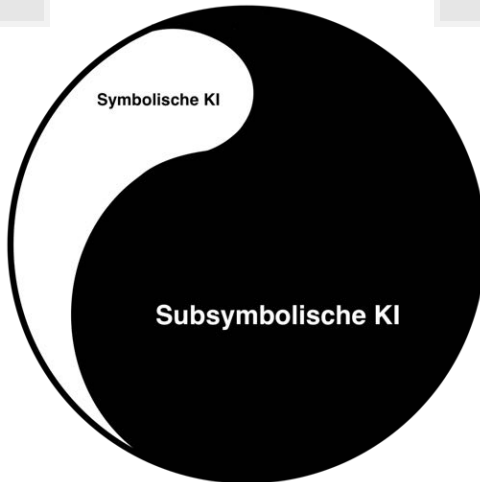
Korrelationen
Muster
Robustheit
Lernen
...



Datengetriebene 'Beweise' in der Mathematik ???

Präzises Schließen
Abstraktion
Kausalität
Domänenwissen
...

Korrelationen
Muster
Robustheit
Lernen
...



Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

11 - Primzahl.

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

11 - Primzahl.

13 - Primzahl.

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

11 - Primzahl.

13 - Primzahl.

... wir brechen hier ab und extrapolieren: Satz gilt!

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

11 - Primzahl.

13 - Primzahl.

... wir brechen hier ab und extrapolieren: Satz gilt!

Lernen von Funktions-Approximationen: kein Allheilmittel!

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

11 - Primzahl.

13 - Primzahl.

... wir brechen hier ab und extrapolieren: Satz gilt!

Lernen von Funktions-Approximationen: kein Allheilmittel!

1. Herausforderung: Primzahl-Eigenschaft entdecken!

Datengetriebene 'Beweise' in der Mathematik ???

Behauptung:

Jede ungerade Zahl (> 1) ist eine Primzahl

Beweis: (empirisch, datengetrieben)

3 - Primzahl.

5 - Primzahl.

7 - Primzahl.

~~9 - Messfehler.~~

11 - Primzahl.

13 - Primzahl.

... wir brechen hier ab und extrapolieren: Satz gilt!

Lernen von Funktions-Approximationen: kein Allheilmittel!

1. Herausforderung: **Primzahl-Eigenschaft entdecken!**

2. Herausforderung: **Primzahl-Eigenschaft erklären!**

Deduktiver Beweis — Mathematik

$A \cup B := \dots$

$A \cap B := \dots$

$A \subseteq B \Leftrightarrow \dots$

$A = B := \dots$

\dots

\dots

Logik-Regeln

Annahmen

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Theorem

Deduktiver Beweis — Mathematik

$$A \cup B := \dots$$

$$A \cap B := \dots$$

$$A \subseteq B \Leftrightarrow \dots$$

$$A = B := \dots$$

...

...

Logik-Regeln

Annahmen

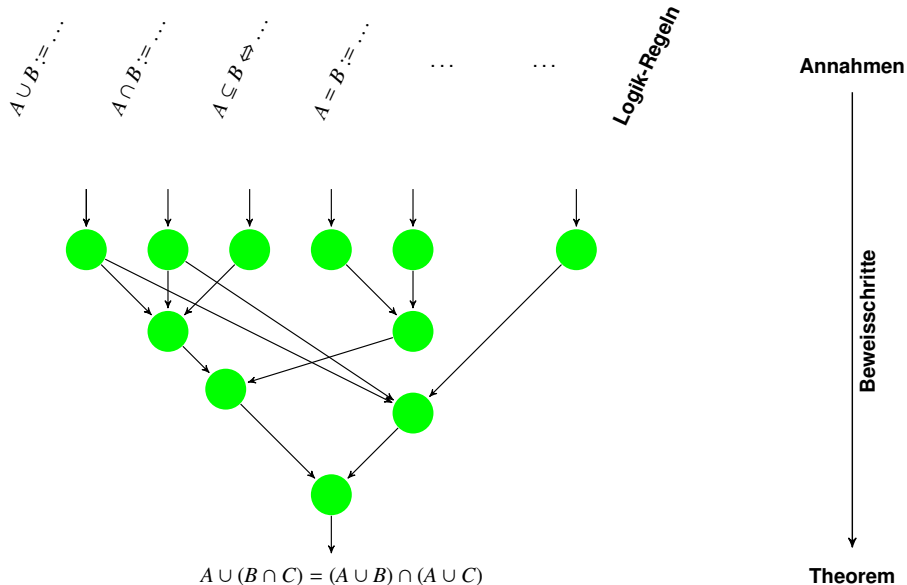
Beweisschritte



Theorem

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Deduktiver Beweis — Mathematik



Deduktiver Beweis — Mathematik

$A \cup B := \dots$

$A \cap B := \dots$

$A \subseteq B \Leftrightarrow \dots$

$A = B := \dots$

...

...

Logik-Regeln

Annahmen

Theorembeweiser:

Computerprogramme ...

... die solche Beweise vollautomatisch suchen

Unser Beweiser LEO-III ist weltweit führend ...

... für (polymorphe) Logik höherer Stufe

Weitere Anwendungen: Verifikation von ...

... Software/Hardware (Informatik)

... Rationale Argumente (Philosophie)

Beweisschritte

Theorem

Deduktiver Beweis — Mathematik

$$A \cup B := \dots$$

$$A \cap B := \dots$$

$$A \subseteq B \Leftrightarrow \dots$$

$$A = B := \dots$$

...

...

Logik-Regeln

Annahmen

Theorembeweiser:

Computerprogramme
... die solche

Unser Beweiser LEO-
... für (polymod)

Weitere Anwendungen
... Software/Hardware
... Rationale Axiome

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's God Theorem

Holy Logic: Computer Scientists 'Prove' God Exists

By David Knight



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.

‘Beweise’ in der Künstlichen Intelligenz



‘Beweise’ in der Künstlichen Intelligenz



Komplexitätssteigerung auf mehreren Ebenen:

komplexe, interagierende traditionelle Software
zunehmender Einsatz von datengetriebener KI

‘Beweise’ in der Künstlichen Intelligenz



‘Beweise’ in der Künstlichen Intelligenz

FEBRUARY 16, 2020

SPACEFLIGHT NOW

HOME NEWS ARCHIVE LAUNCH SCHEDULE MISSION REPORTS SUBSCRIBE MEMBERSHIP

BREAKING NEWS > [February 14, 2020] SpaceX test-fires Falcon 9 rocket; Starlink launch delayed to Monday

NASA, Boeing managers admit problems with Starliner software verification

February 7, 2020 William Harwood



Artist's concept of the Starliner service module (top) separating from the Starliner crew module before re-entry. Credit: Boeing

Two software errors detected after launch of a Boeing Starliner crew ship during an unplanned test flight last December, one of which prevented a planned docking with the International Space Station, could have led to catastrophic failures had they not been caught and corrected in time, NASA said Friday.



‘Beweise’ in der Künstlichen Intelligenz



Wenn es schief geht brauchen wir:

präzise Erklärungen und Identifikation von Kausalitäten
verlässliche Klärung von Verantwortung und Haftung

'Beweise' in der Künstlichen Intelligenz



‘Beweise’ in der Künstlichen Intelligenz

Superintelligenz:

Superintelligenz noch nicht in Sicht

Überzogene Erwartungen an datengetriebene KI

Allerdings: sehr hohe Dynamik im Gebiet



‘Beweise’ in der Künstlichen Intelligenz

Superintelligenz:

Superintelligenz noch nicht in Sicht

Überzogene Erwartungen an datengetriebene KI

Allerdings: sehr hohe Dynamik im Gebiet

Gerade deshalb sind die Herausforderungen groß:

Die unreflektierte Anwendung “unreifer” KI-Technologie in kritischen Anwendungsgebieten ist das Problem

‘Beweise’ in der Künstlichen Intelligenz

Superintelligenz:

Superintelligenz noch nicht in Sicht

Überzogene Erwartungen an datengetriebene KI

Allerdings: sehr hohe Dynamik im Gebiet

Gerade deshalb sind die Herausforderungen groß:

Die unreflektierte Anwendung “unreifer” KI-Technologie in kritischen Anwendungsgebieten ist das Problem

Was wir brauchen (in kritischen Anwendungen):

Deduktive Kontrolle von ethischen-rechtlichen Vorgaben
weil: “Erlernen” solcher Vorgaben ist problematisch

Erhaltung von Leben — Eine moralische Maxime von Maschinen?

Pilot über die Boeing 737 Max

SPIEGEL 

"Eine Automatisierung will nicht überleben. Wir schon"



Uwe Harter ist seit 26 Jahren Pilot von Passagierflugzeugen. Er steuert A320-Jets - das Pendant von Airbus zur Boeing 737. Ein Gespräch über Notfälle im Cockpit und die Schulung der Crew. Von Claus Hecking **mehr...**

737 Max: FBI schließt sich offenbar Ermittlungen wegen Zulassung an
Abstürze der Boeing 737 Max: Welche Rolle spielten die Piloten?

- ▶ Können KI Systeme eine *eigene Ethik* entwickeln? —**Ich bezweifle das!**—
- ▶ Können KI System durch *unsere ethischen Prinzipien* kontrolliert werden? —**Möglicherweise, aber das ist nicht einfach!**—

Erhaltung von Leben — Eine moralische Maxime von Maschinen?

Pilot über die Boeing 737 Max

SPIEGEL 

"Eine Automatisierung will nicht überleben. Wir schon"



Uwe Harter ist seit 26 Jahren Pilot von Passagierflugzeugen. Er steuert A320-Jets - das Pendant von Airbus zur Boeing 737. Ein Gespräch über Notfälle im Cockpit und die Schulung der Crew. Von Claus Hecking **mehr...**

737 Max: FBI schließt sich offenbar Ermittlungen wegen Zulassung an
Abstürze der Boeing 737 Max: Welche Rolle spielten die Piloten?

- ▶ Können KI Systeme eine *eigene Ethik* entwickeln? —**Ich bezweifle das!**—
- ▶ Können KI System durch *unsere ethischen Prinzipien* kontrolliert werden? —**Möglicherweise, aber das ist nicht einfach!**—

Moral Machine Experiment (siehe Nature, vol. 563)

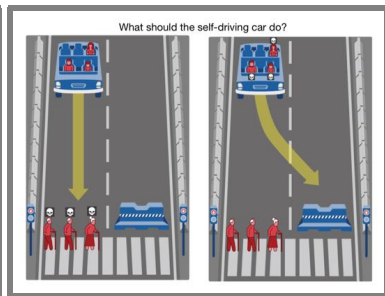
ARTICLE

<https://doi.org/10.1038/441586-018-0637-6>

The Moral Machine experiment

Edmond Awad¹, Sohan Dsouza¹, Richard Kim¹, Jonathan Schulz², Joseph Henrich², Azim Shariff^{3*}, Jean-François Bonnefon^{4*} & Iyad Rahwan⁵⁻⁶

With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behaviour. To address this challenge, we deployed the Moral Machine, an online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories. Here we describe the results of this experiment. First, we summarize global moral preferences. Second, we document individual variations in preferences, based on respondents' demographics. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries. Fourth, we show that these differences correlate with modern institutions and deep cultural traits. We discuss how these preferences can contribute to developing global, socially acceptable principles for machine ethics. All data used in this article are publicly available.



Präferenzen — teilweise kultur-/kontextabhängig — einige Beispiele:

Global: Menschen vor Tieren, mehr-Leben vor weniger-Leben

Kultur: 'jünger vor älter' weniger stark ausgeprägt in Asien

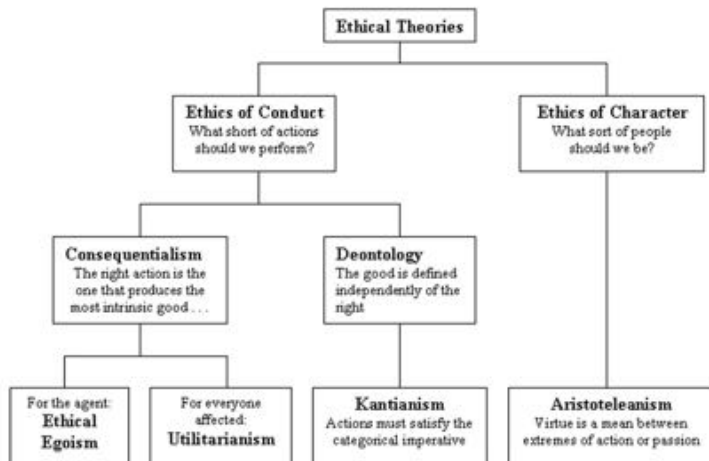
Länder: 'Status vor kein-Status' höher ausgeprägt in reichen Ländern

Teilweise im Widerspruch zu den Empfehlungen in:

C. Luetge, **The German Ethics Code for automated and connected driving.**

Philos. Technol. 30, 547–558 (2017).

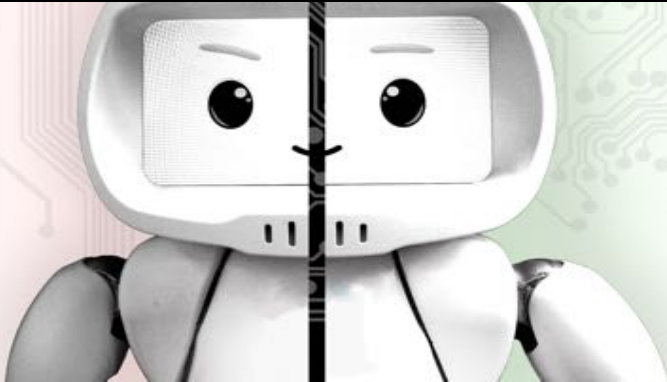
Welche Ethik?



Ethics

Koexistenz mit **Intelligenten Autonomen Systemen (IASs)**?

- ▶ geeignete **Kontrollmechanismen** für IASs
- ▶ geeignete Form der **Mensch-Maschinen-Interaktion**



Ethics

Koexistenz mit **Intelligenten Autonomen Systemen (IASs)**?

- ▶ geeignete **Kontrollmechanismen** für IASs
- ▶ geeignete Form der **Mensch-Maschinen-Interaktion**

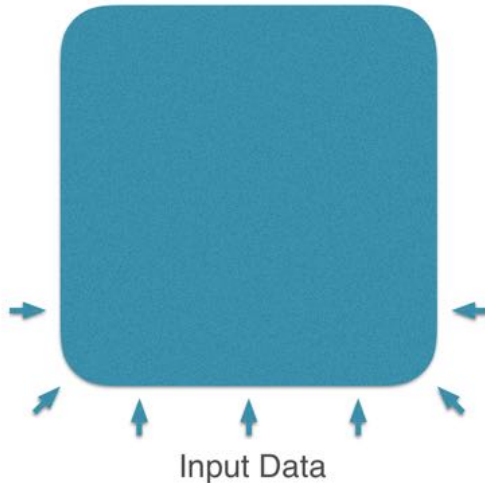
Existierende gesellschaftliche Prozesse basieren auf:

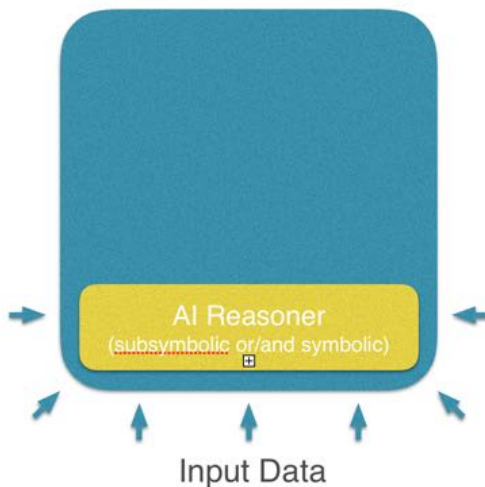
- ▶ **Erklärungen, rationaler Argumentation & Dialog,**
- ▶ inklusive **explizitem normativem Schließen** (rechtlich & ethisch)

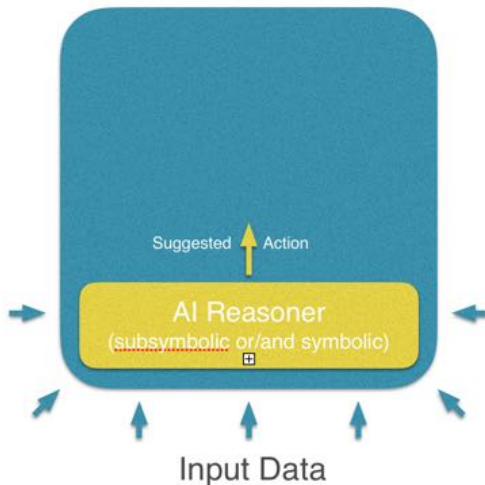
Entwicklung von IASs ohne solche Kompetenzen? Wie sinnvoll?

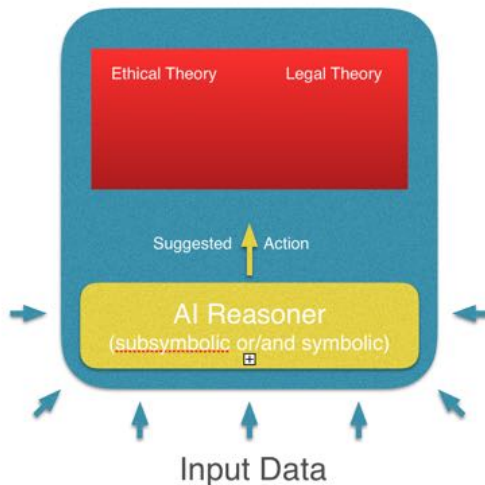


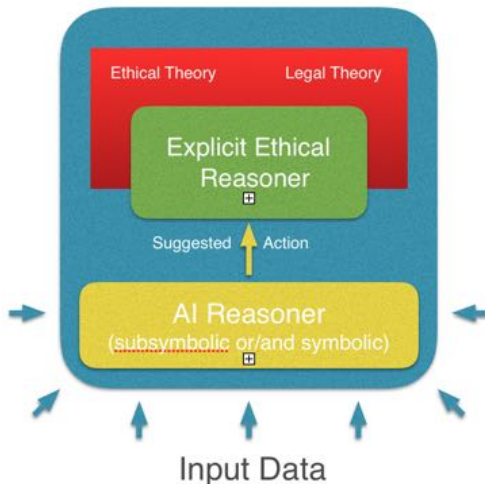
IAS

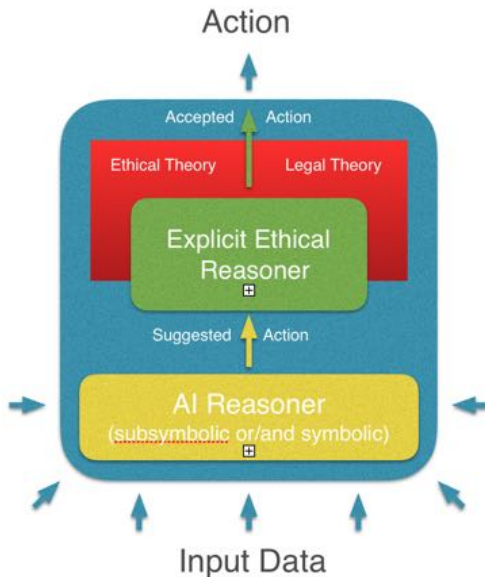


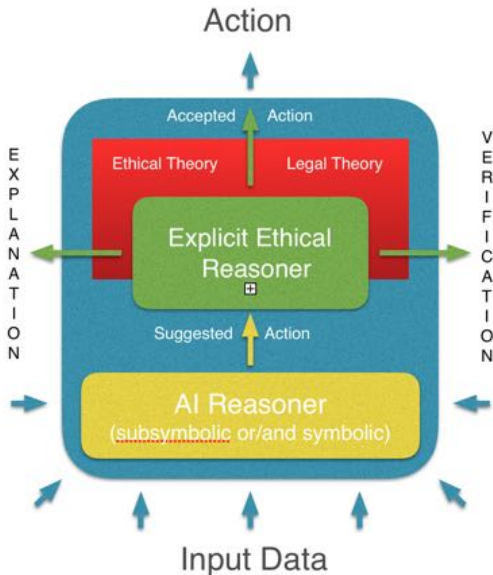


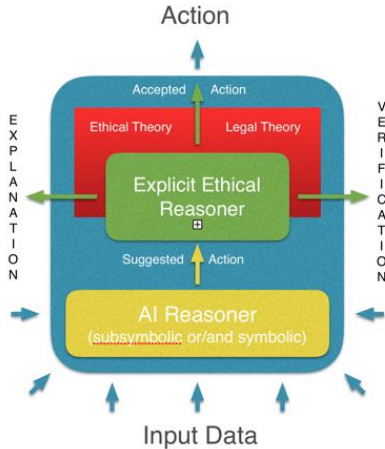












Verwandte Arbeiten

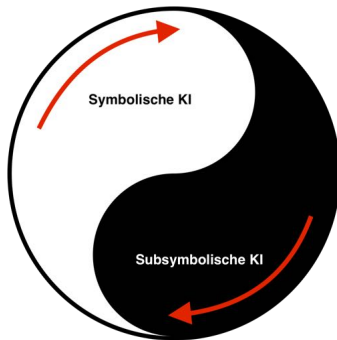
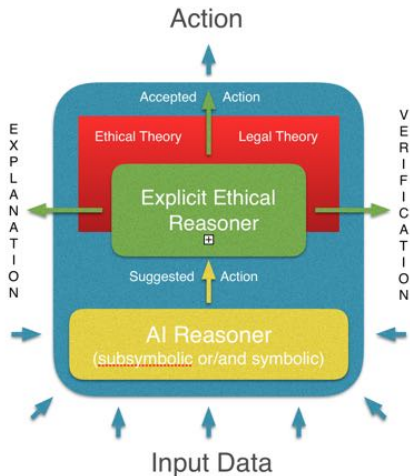
- ▶ Toward Ethical Robots
 - ▶ [ArkoudasEtAl., 2005]
- ▶ Artificial Moral Agents
 - ▶ [Wallach&Allen, 2008]
- ▶ Ethical Governors
 - ▶ [ArkinEtAl., 2009, 2012]
 - ▶ [Dennis&Fisher, 2017]
- ▶ Ethical Deliberation in ART
 - ▶ [Dignum, 2017]
- ▶ Programming Machine Ethics
 - ▶ [Pereira&Saptawijaya, 2016]

Adressiert Forderungen nach Transparenz, Erklärbarkeit, Verifizierbarkeit:

- *"Ethics Guidelines for Trustworthy AI"* [HLEG, EC, April 2019]
- *"Policy and Investment Recommend. for Trustworthy AI"* [HLEG, EC, June 2019]
- *"Strategie Künstliche Intelligenz"* [Bundesregierung, November 2018]

Pseudo-Ethischer KI Agent

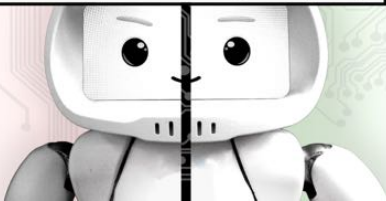
Trustworthy AI



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)



— Universelles Logisches Schließen —

Ontologisches Argument für die Existenz Gottes



Kurt Gödel (1914-1976) mit Einstein

Definition:

Eine *Gott-artige Entität* besitzt alle positive Eigenschaften.

(plus weitere Axioms und Definitionen)

Theorem:

Notwendigerweise existiert Gott.

Anselm v. C.
Gaunilo

Th. Aquinas

Descartes
Spinoza
Leibniz

Hume
Kant

Hegel

Frege

Hartshorne
Malcolm
Lewis
Plantinga
Gödel

Computational Metaphysics — Gödel's Ontological Argument (1970)

Ontologische Beweise Feb. 10, 1970

$\mathcal{P}(\varphi)$ φ is positive (i.e. $\varphi \in \mathcal{P}$)
Axiom 1: $\mathcal{P}(\varphi) \cdot \mathcal{P}(\psi) \supset \mathcal{P}(\varphi \cdot \psi)$ Axiom 2: $\mathcal{P}(\varphi) \supset \mathcal{P}(\Box \varphi)$
[1] $G(x) = (\varphi) [\mathcal{P}(\varphi) \supset \varphi(x)]$ (Gödel)
[2] $\varphi \text{ Ess } x = (\psi) [\psi(x) \supset (\forall \varphi) (\mathcal{P}(\varphi) \supset \varphi(x))]$ (Essence of x)
 $\mathcal{P} \supset \mathcal{N} = N(\mathcal{P} \supset \mathcal{Q})$ Necessity
Axiom 2: $\mathcal{P}(\varphi) \supset N\mathcal{P}(\varphi)$
 $\neg \mathcal{P}(\varphi) \supset N \neg \mathcal{P}(\varphi)$ } because it follows from the nature of the

$M(x) G(x)$ means: all possible This is
Axiom 4: $\mathcal{P}(\varphi) \cdot \varphi \supset \Box \varphi$
Lower $\{ x = x \}$ is possible
upper $\{ x \neq x \}$ is impossible
But if a system S of \mathcal{P} would mean that S is impossible) would be $x \neq x$



Herausforderungen (Ontologisches Argument):

- ▶ Modalitäten: "Möglicherweise gilt φ " vs. "Notwendigerweise gilt φ "
- ▶ Unterschiedliche erststufige und höherstufige Quantoren
- ▶ Welche Logik/Logikkombination?

Herausforderungen (Ethische Theorien und Recht):

- ▶ Modalitäten: " φ ist verboten" vs. " φ ist erlaubt"
- ▶ Unterschiedliche erststufige und höherstufige Quantoren
- ▶ Welche Logik/Logikkombination?

Ambitioniertes Forschungsspektrum

Von der Formalen Analyse und Verifikation von

- ▶ Varianten des Ontologischen Gottesbeweises
- ▶ Grundlagen der Metaphysik (z.B. Principia Logico-Metaphysica)
- ▶ Grundlagen der Mathematik (z.B. Category Theory)

über

- ▶ Rationalen Argumenten in Politik, Recht und Ethik, ...

zu

- ▶ Ethisch-Rechtlicher Kontrolle von Intelligenten Autonomen Systemen
- ▶ Trustworthy AI made in Europe

Ambitioniertes Forschungsspektrum

Von der Formalen Analyse und Verifikation von

- ▶ Varianten des Ontologischen Gottesbeweises
- ▶ Grundlagen der Metaphysik (z.B. Principia Logico-Metaphysica)
- ▶ Grundlagen der Mathematik (z.B. Category Theory)

über

- ▶ Rationalen Argumenten in Politik, Recht und Ethik, ...

zu

- ▶ Ethisch-Rechtlicher Kontrolle von Intelligenten Autonomen Systemen
- ▶ Trustworthy AI made in Europe

Technologie: Universelles Logisches Schließen

Ambitioniertes Forschungsspektrum

Von der Formalen Analyse und Verifikation von

- ▶ Varianten des Ontologischen Gottesbeweises
- ▶ Grundlagen der Metaphysik (z.B. Principia Logico-Metaphysica)
- ▶ Grundlagen der Mathematik (z.B. Category Theory)

über

- ▶ Rationalen Argumenten in Politik, Recht und Ethik, ...

zu

- ▶ Ethisch-Rechtlicher Kontrolle von Intelligenten Autonomen Systemen
- ▶ Trustworthy AI made in Europe

Technologie: Universelles Logisches Schließen

“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)

Wissensrepräsentation & Schließen: Universeller Ansatz



Eine Geschichte über die Bändigung des Logik-Zoos



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

Wissensrepräsentation & Schließen: Universeller Ansatz

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 tacitly assume that $PROP$ is countably 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

Wissensrepräsentation & Schließen: Universeller Ansatz

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 $\mathfrak{M}, v \models \varphi$.

Wissensrepräsentation & Schließen: Universeller Ansatz

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 1

$V(p)$

$(W, \{$

in the

Metalanguage

in a model is a binary relation on W ,

position symbol p in PROP a subset

p is true. The first two components

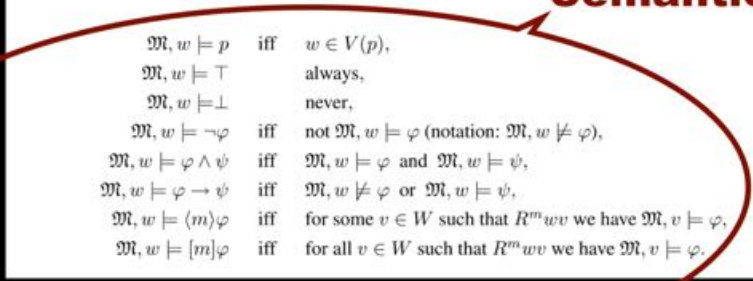
: model. If there is only one relation

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

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 $\mathfrak{M}, v \models \varphi$.

Klassische Höherstufige Logik (HOL) als Universelle (Meta-)Logik

HOL (Meta-Logik)

$\varphi ::=$ 

Deine Logik (Objekt-Logic)

$\psi ::=$ 

Einbettung  in 

 = 

 = 

 = 

 = 

Einbettung meta-logischer Begriffe:  in 

gültig = 

erfüllbar = 

... = 

Diese Menge von Gleichungen wird an Theorembeweiser übergeben
(als Hintergrundtheorie)

Klassische Höherstufige Logik (HOL) als Universelle (Meta-)Logik

HOL $s, t ::= c_\alpha \mid x_\alpha \mid (\lambda x_\alpha s_\beta)_{\alpha \rightarrow \beta} \mid (s_{\alpha \rightarrow \beta} t_\alpha)_\beta \mid \neg s_o \mid s_o \vee t_o \mid \forall x_\alpha t_o$

HOML $\varphi, \psi ::= \dots \mid \neg \varphi \mid \varphi \wedge \psi \mid \varphi \rightarrow \psi \mid \Box \varphi \mid \Diamond \varphi \mid \forall x_\gamma \varphi \mid \exists x_\gamma \varphi$

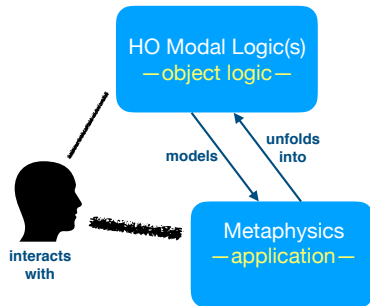
\neg	$=$	$\lambda \varphi_{\mu \rightarrow o} \lambda w_\mu \neg \varphi w$
\wedge	$=$	$\lambda \varphi_{\mu \rightarrow o} \lambda \psi_{\mu \rightarrow o} \lambda w_\mu (\varphi w \wedge \psi w)$
\rightarrow	$=$	$\lambda \varphi_{\mu \rightarrow o} \lambda \psi_{\mu \rightarrow o} \lambda w_\mu (\neg \varphi w \vee \psi w)$
\forall	$=$	$\lambda h_{\gamma \rightarrow (\mu \rightarrow o)} \lambda w_\mu \forall d_\gamma h d w$
\exists	$=$	$\lambda h_{\gamma \rightarrow (\mu \rightarrow o)} \lambda w_\mu \exists d_\gamma h d w$
\Box	$=$	$\lambda \varphi_{\mu \rightarrow o} \lambda w_\mu \forall u_\mu (\neg r w u \vee \varphi u)$
\Diamond	$=$	$\lambda \varphi_{\mu \rightarrow o} \lambda w_\mu \exists u_\mu (r w u \wedge \varphi u)$
gültig	$=$	$\lambda \varphi_{\mu \rightarrow o} \forall w_\mu \varphi w$

Diese Menge von Gleichungen wird an Theorembeweiser übergeben
(als Hintergrundtheorie)

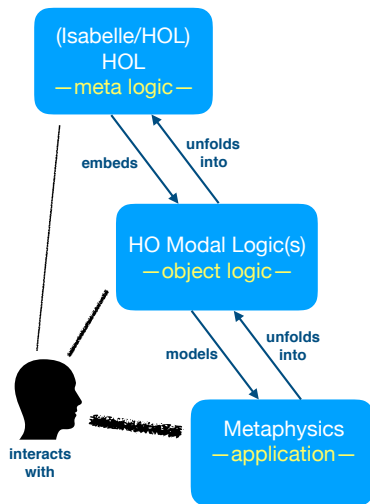
Universelles Logisches Schließen in Isabelle/HOL



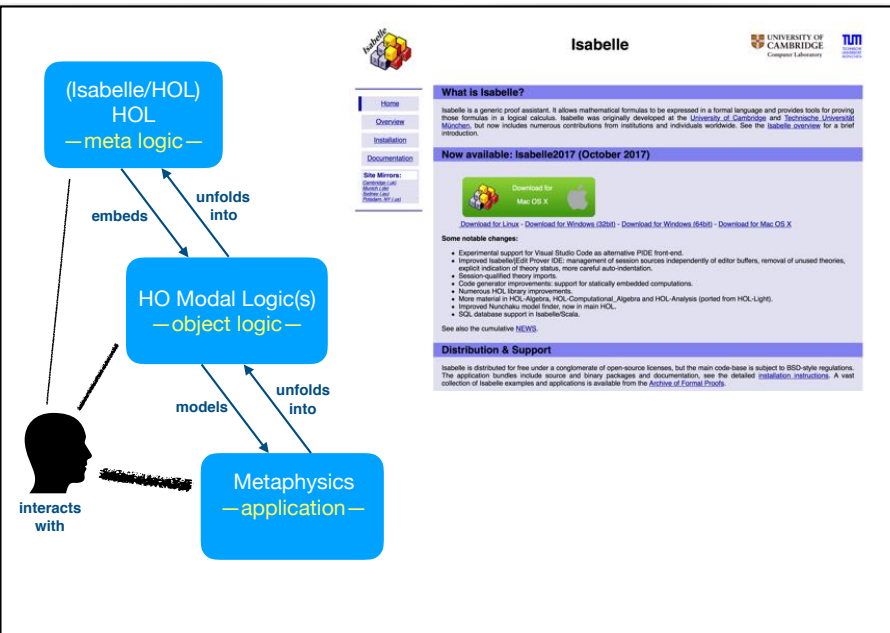
Universelles Logisches Schließen in Isabelle/HOL



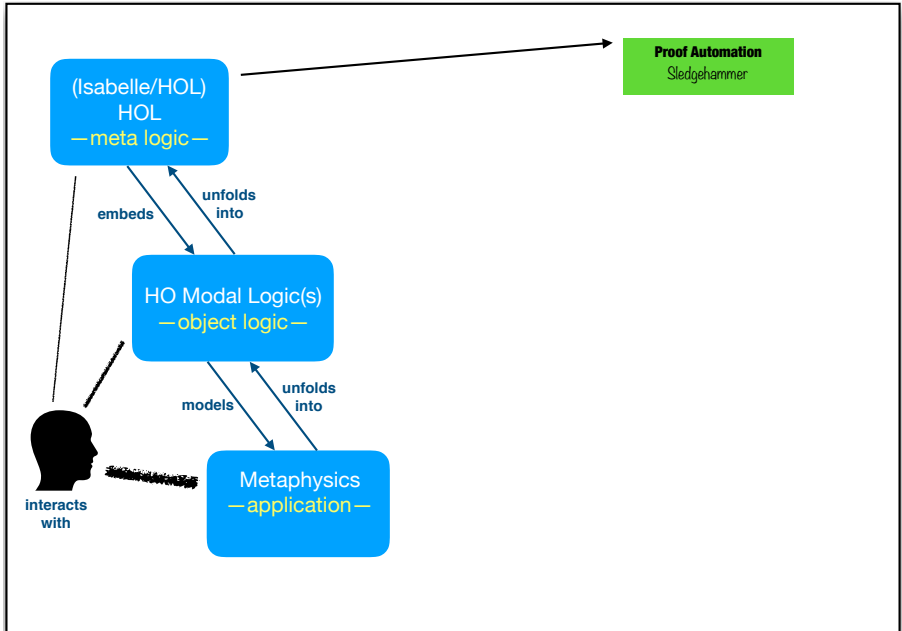
Universelles Logisches Schließen in Isabelle/HOL



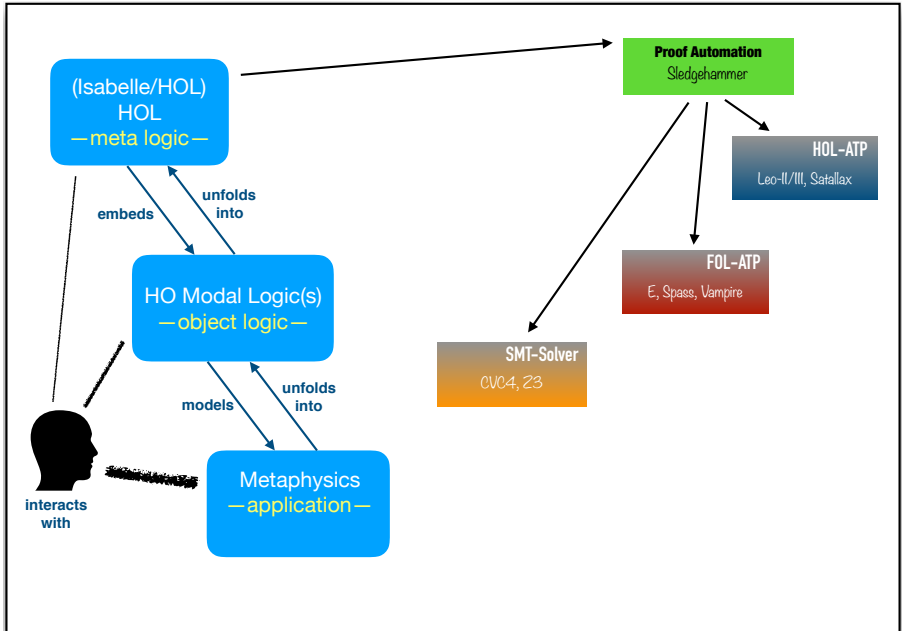
Universelles Logisches Schließen in Isabelle/HOL



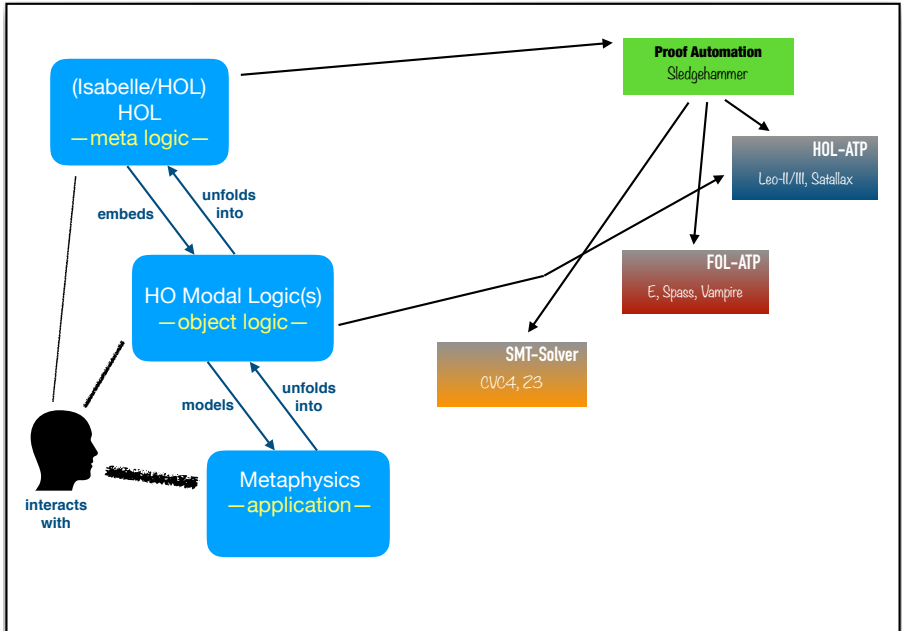
Universelles Logisches Schließen in Isabelle/HOL



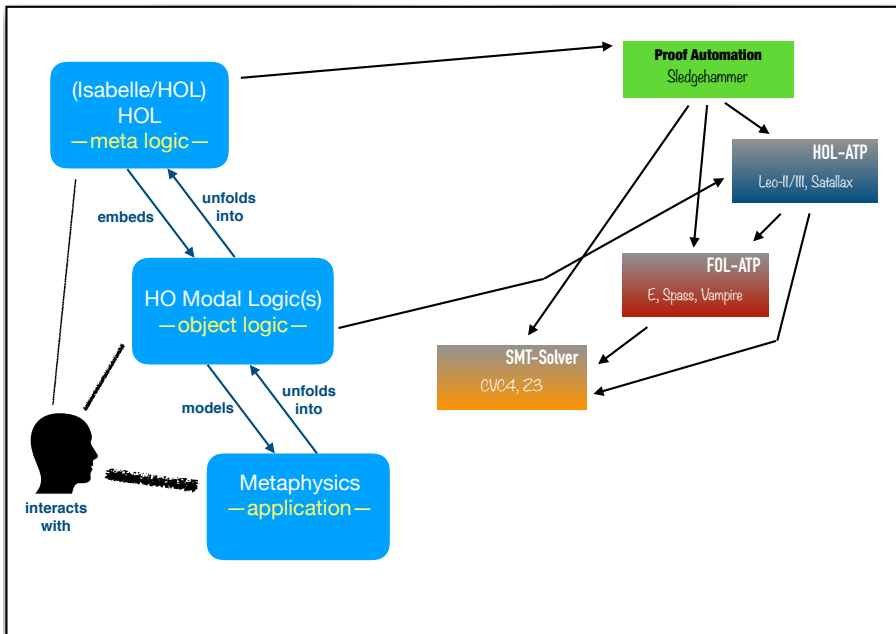
Universelles Logisches Schließen in Isabelle/HOL



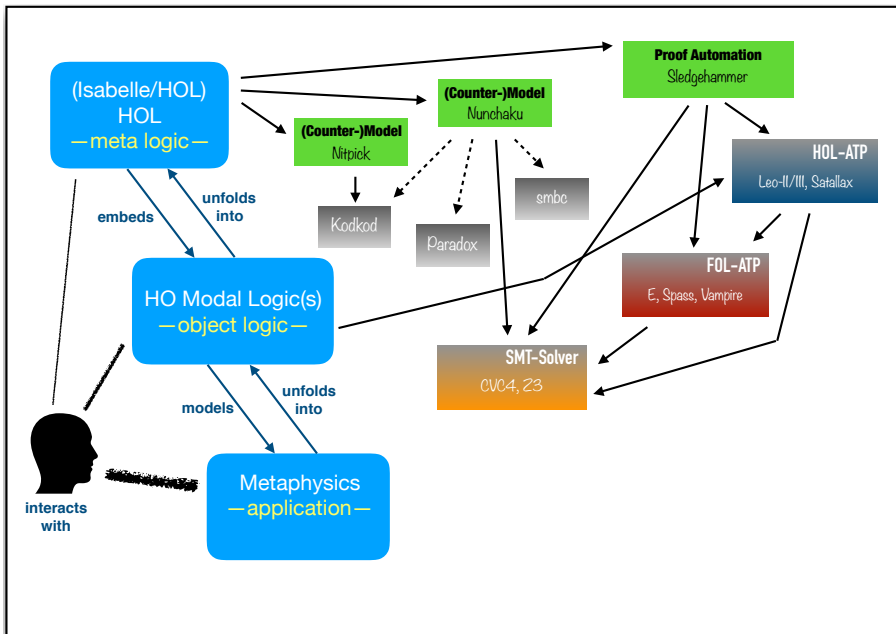
Universelles Logisches Schließen in Isabelle/HOL



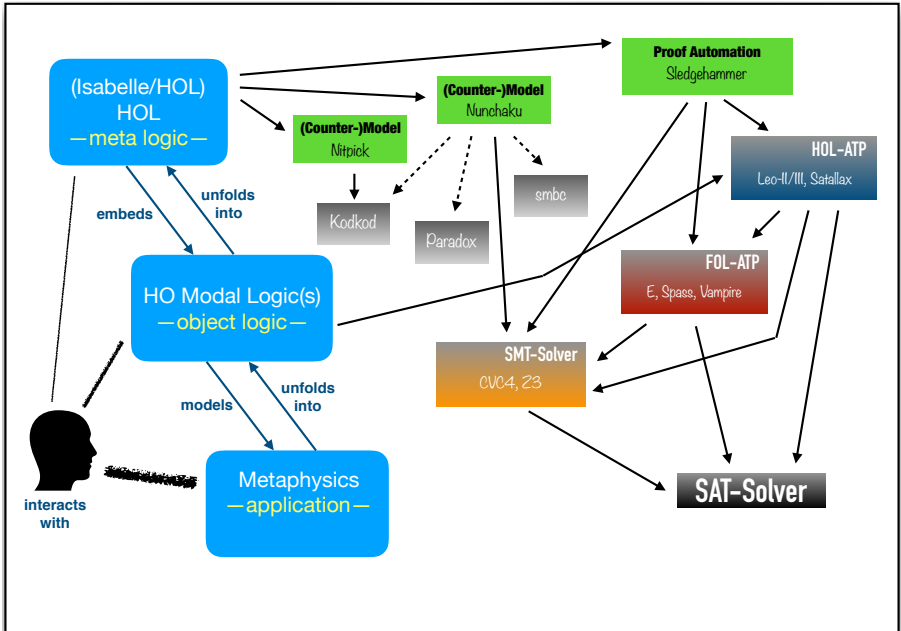
Universelles Logisches Schließen in Isabelle/HOL



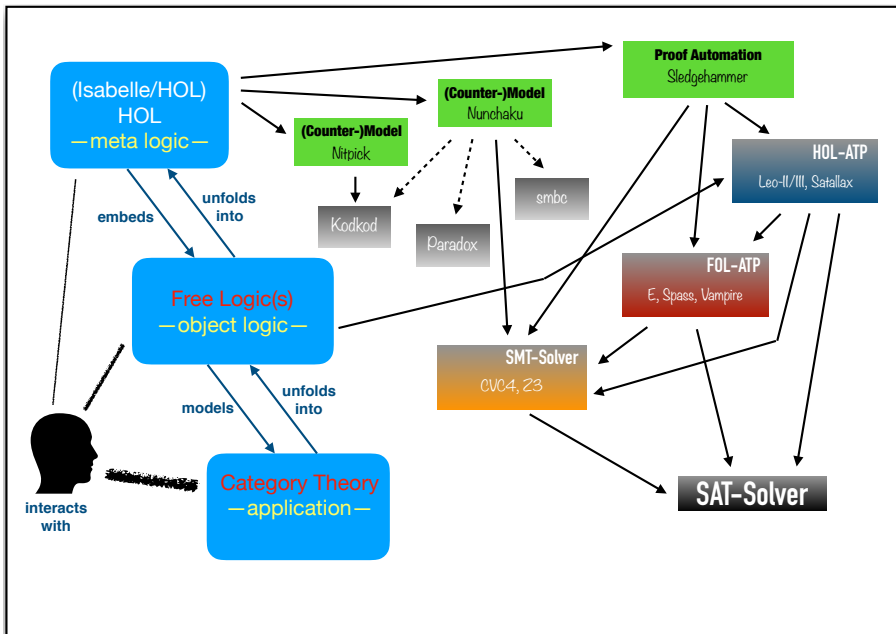
Universelles Logisches Schließen in Isabelle/HOL



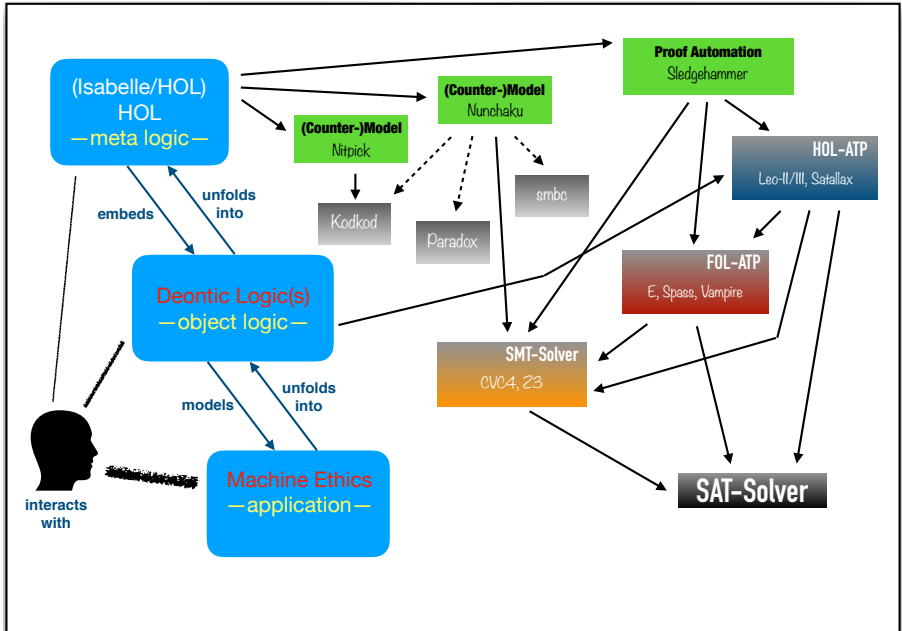
Universelles Logisches Schließen in Isabelle/HOL



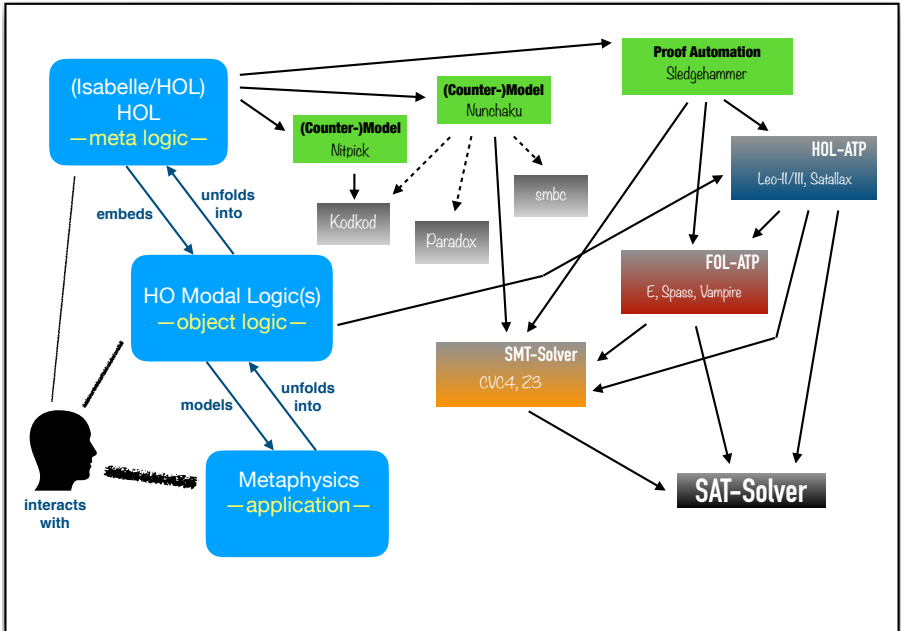
Universelles Logisches Schließen in Isabelle/HOL



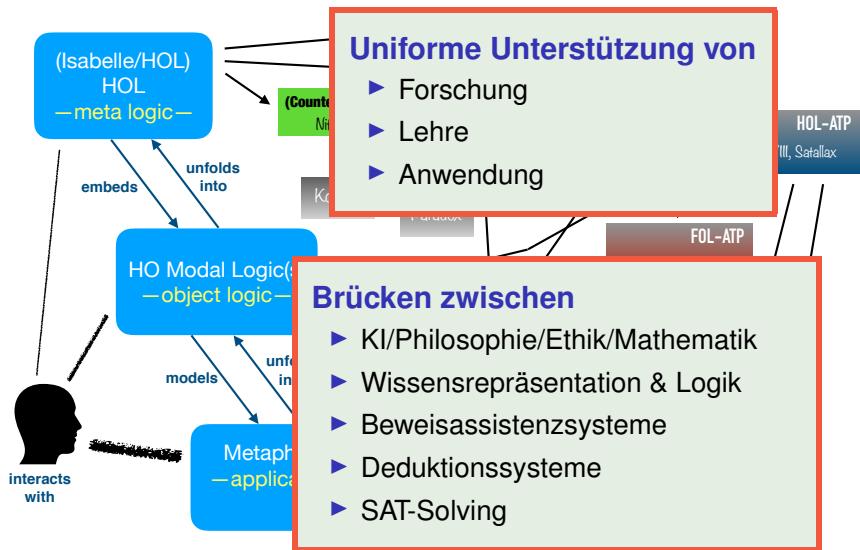
Universelles Logisches Schließen in Isabelle/HOL



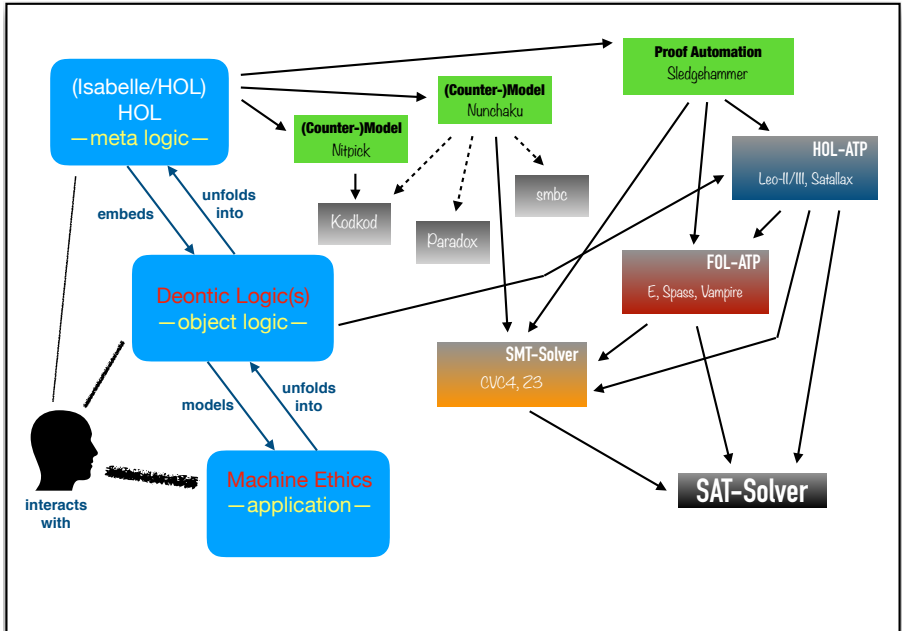
Universelles Logisches Schließen in Isabelle/HOL

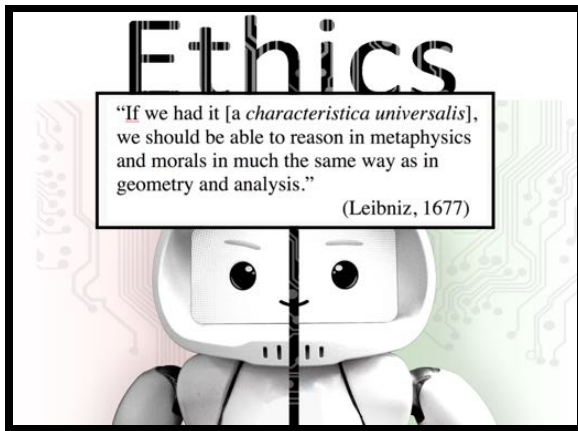


Universelles Logisches Schließen in Isabelle/HOL



Universelles Logisches Schließen in Isabelle/HOL





— Normatives Schließen in Ethisch-Rechtlichen Theorien —

Normatives Schließen

Herausforderungen: Welche Ethik? Welches Recht? **Welche Logik(en)?**

- ▶ Geeignete Modellierung des Begriffs der **Obligation**
- ▶ **Obligation:** nicht-trivialer \square -Operator der Modallogik/Deontischen Logik
- ▶ Problem: “Contrary-to-duty” (**CTD**) Szenarien

Standard CTD Struktur (Chisholm)

1. obligatorisch '*a*'
2. obligatorisch 'falls *a* dann *b*'
3. wenn 'nicht *a*' dann obligatorisch '*b*'
4. 'nicht *a*' (in gegebener Situation)

Gefahr: Paradoxie/Inkonsistenz— Ex falso quodlibet!

Normatives Schließen

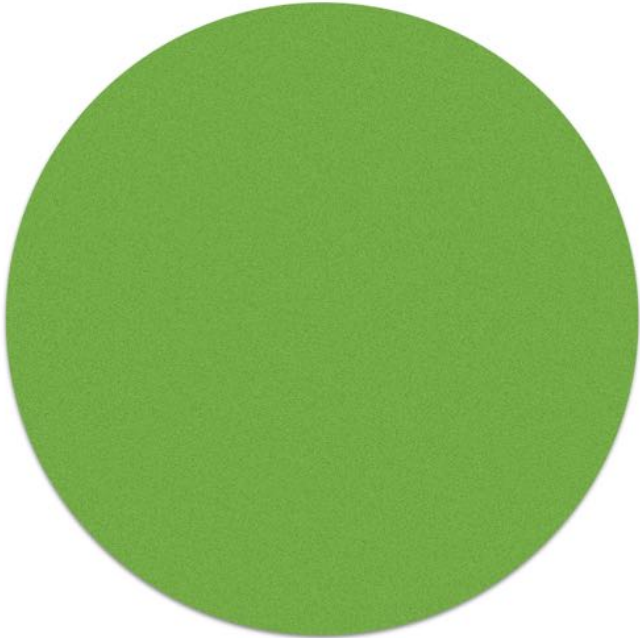
Herausforderungen: Welche Ethik? Welches Recht? **Welche Logik(en)?**

- ▶ Geeignete Modellierung des Begriffs der **Obligation**
- ▶ **Obligation:** nicht-trivialer \square -Operator der Modallogik/Deontischen Logik
- ▶ Problem: “Contrary-to-duty” (**CTD**) Szenarien

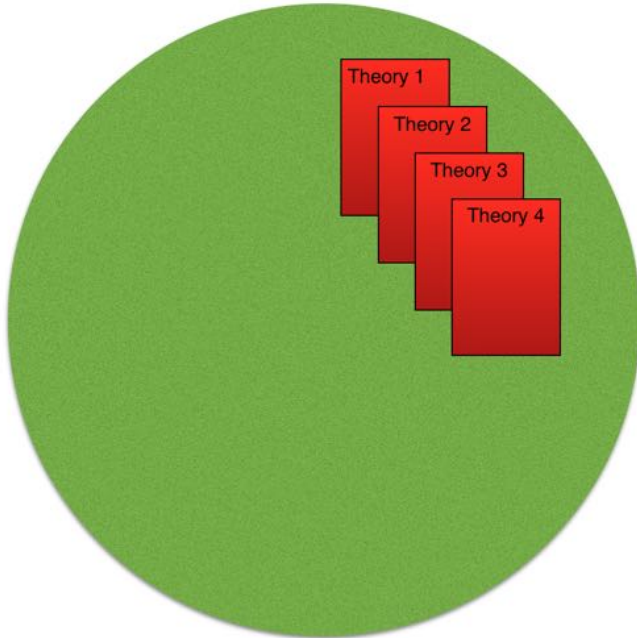
CTD Beispiel (X. Parent): **EU GDPR**

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. **Implizit:** 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. **Gegebene Situation:** Some personal data has been processed unlawfully.

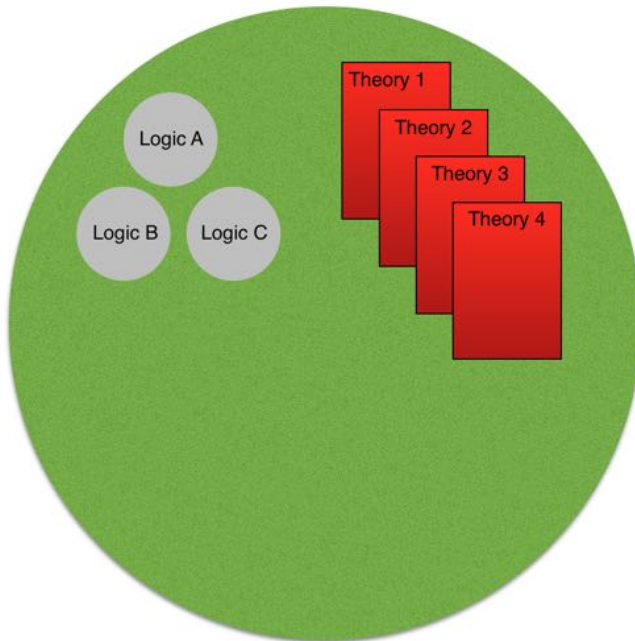
Gefahr: Paradoxie/Inkonsistenz — Ex falso quodlibet!



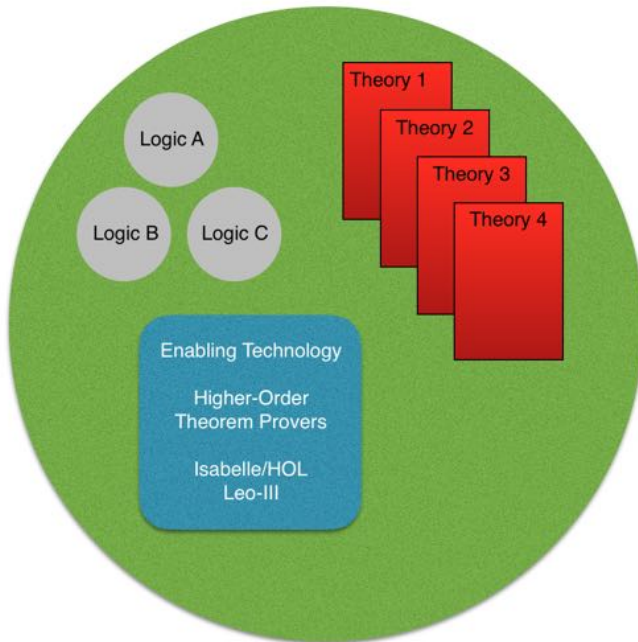
Experimentierplattform für Normatives Schließen und Maschinen-Ethik



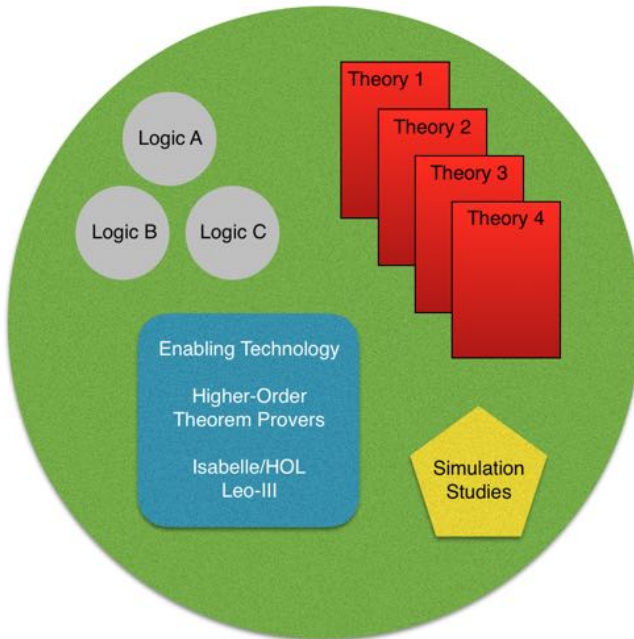
Experimentierplattform für Normatives Schließen und Maschinen-Ethik



Experimentierplattform für Normatives Schließen und Maschinen-Ethik



Experimentierplattform für Normatives Schließen und Maschinen-Ethik



Experimentierplattform für Normatives Schließen — Demo möglich!

The screenshot displays the Isabelle/HOL Sledgehammer interface. The top pane shows a theory file named `GDPR.thy` with the following content:

```
1 theory GDPR imports SDL (* Christoph Benzmüller & Xavier Parent, 2018 *)
2
3 begin (** GDPR Example **)
4 consts process_data_lawfully:: $\sigma$  erase_data:: $\sigma$  kill_boss:: $\sigma$ 
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  $\rightarrow$   $\neg$ erase_data)]" and
11  (* If data was not processed lawfully, then it is an obligation to erase the data. *)
12  A2: "[ $\neg$ process_data_lawfully  $\rightarrow$  0(erase_data)]"
13  (* Given a situation where data is processed unlawfully. *) and
14  A3: "[ $\neg$ process_data_lawfully]v"
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( $\neg$ 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 bottom pane shows the Sledgehammer output:

```
Sledgehammering...
Proof found...
"spass": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could
"e": The prover derived "False" from "A1", "A2", "A3", "D", and "Implicit", which could be d
"cv4": Try this: by (metis A1 A2 A3 D Implicit) (68 ms)
"z3": Try this: by (metis A1 A2 A3 D Implicit) (59 ms)
```

The interface includes a toolbar at the top, a sidebar on the right with links to Documentation, Sidekick, State, and Theories, and a bottom bar with tabs for Output, Query, Sledgehammer, and Symbols.

Isabelle/HOL (Beweisassistent) als Werkzeug für Universelles Schließen

Experimentierplattform für Normatives Schließen — Demo möglich!

```
1 theory GDPR imports SDL                                (* Christoph Benzmüller & Xavier Parent, 2018 *)
2
3 begin (** GDPR Example **)
4   consts process_data_lawfully:: $\sigma$  erase_data:: $\sigma$  kill_boss:: $\sigma$ 
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  $\rightarrow$   $\neg$ erase_data)]" and
11    (* If data was not processed lawfully, then it is an obligation to erase the data. *)
12    A2: "[ $\neg$ process_data_lawfully  $\rightarrow$  0(erase_data)]"
13    (* Given a situation where data is processed unlawfully. *) and
14    A3: "[ $\neg$ process_data_lawfully]iv"
15
16  (** Some Experiments **)
17  lemma True nitpick [satisfy] oncs (* Consistency-check: Is there a model? *)
18  |
19  |
20  |
21  |
22  |
23 end
```

Gefahren-Zone:
Paradoxien and Inkonsistenzen!

✓ Proof state ✓ Auto update Update Search: 100%

Output Query Sledgehammer Symbols

Parallele Entwicklung und Verifikation von
ethisch-rechtlichen Theorien \leftrightarrow passende Logikformalismen

Isabelle/HOL (Beweisassistent) als Werkzeug für Universelles Schließen

Formalisierung einer Ethischen Theorie: Alan Gewirth's "PGC"

"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)

Abwandlung & Erweiterung der **Goldenen Regel**:

"Behandle andere so, wie du von ihnen behandelt werden willst."

"Was du nicht willst, dass man dir tu', das füg auch keinem andern zu."

Referenzen

- ▶ A. Gewirth. Reason and morality. U of Chicago Press, 1978. (401 pages)
- ▶ 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. (523 pages)
- ▶ A. Kornai. Bounding the impact of AGI. J. Experimental & Theoretical AI, 2014.

Formalisierung einer Ethischen Theorie: Alan Gewirth's "PGC"

"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)

Abwandlung & Erweiterung der **Goldenen Regel**:

"Behandle andere so, wie du von ihnen behandelt werden willst."

"Was du nicht willst, dass man dir tu', das füg auch keinem andern zu."



FORMALISATION AND EVALUATION OF ALAN GEWIRTH'S PROOF FOR THE PRINCIPLE OF GENERIC CONSISTENCY IN ISABELLE/HOL

Title:

Formalisation and Evaluation of Alan Gewirth's Proof for the Principle of Generic Consistency in Isabelle/HOL

Authors:

David Fuenmayor (davfuenmayor /at/ gmail /dot/ com) and [Christoph Benz Müller](#)

[Home](#)

Formalisierung einer Ethischen Theorie: Alan Gewirth's "PGC"

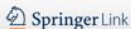
"Act in accord with the generic rights of your recipients as well as of yourself. I shall call this the **Principle of Generic Consistency** (PGC), since it combines the formal consideration of consistency with the material consideration of rights to the generic features or goods of action."
(Alan Gewirth,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

Abwandlung & Erweiterung der **Goldenen Regel**:

*"Behandle andere so, wie du von ihnen behandelt werden willst."
"Was du nicht willst, dass man dir tu", das füg auch keinem andern zu."*



FORMALISATION AND EVALUATION OF ALAN GEWIRTH'S PROOF FOR THE PRINCIPLE OF GENERIC CONSISTENCY IN ISABELLE/HOL



[Pacific Rim International Conference on Artificial Intelligence](#)

PRICAI 2019: [PRICAI 2019: Trends in Artificial Intelligence](#) pp 418-432 | [Cite as](#)

Harnessing Higher-Order (Meta-)Logic to Represent and Reason with Complex Ethical Theories

Authors

Authors and affiliations

David Fuenmayor , Christoph Benzmüller

Formalisierung einer Ethischen Theorie: Alan Gewirth's "PGC"

"Act in accord
call this the **Pr**
consideration
features or god

f. I shall
e formal
generic
y, 1978)

A
"Behandl
"Was du nic

llst."
ern zu."



[Home](#)

BACHELOR'S THESIS

Modelling the US Constitution in HOL

On establishing a dictatorship with Gödel

Valeria Zahoransky

supervised and examined by
Prof. Dr. Christoph BENZMÜLLER

examined by
Prof. Dr. Jan VON PLATO
(University of Helsinki)

FOR

Formalisierung einer Ethischen Theorie: Alan Gewirth's "PGC"

"Act in accord
call this the P
consideration
features or good

"I shall
e formal
generic
y, 1978)

BACHELOR'S THESIS

Modelling the US Constitution in HOL

On establishing a dictatorship with Gödel

A
"Behandl
"Was du nic

llst."
ern zu."

Conference Paper

Full-text available

Computer-supported Analysis of Arguments in Climate Engineering

January 2020

Conference: 3rd International Conference on Logic and Argumentation (CLAR 2020) - At: 6-9 April
2020, Hangzhou, China

Project: Universal Logical Reasoning



David Fuenmayor ·



Christoph Benzmüller

Home

PROF. DR. JARI VON KLEIN
(University of Helsinki)

Ethics

Projektvorschlag ENoRME



- ▶ **“Explicit Normative Reasoning and Machine-Ethics”**
 - Cloud-basierte Plattform für Universelles (Normatives) Schließen
 - Experimente in der Maschinen-Ethik
- ▶ **Interdisziplinär:** Deduktion, Theorembeweisen, Wissensrepräsentation, Maschinelles Lernen, Formale Methoden, Formale Ethik/Philosophie, Autonome Fahrzeuge, Soziale Robotik, usw.
- ▶ **International** und **Industriebeteiligungen**

Ethics

Zusammenfassung und Fazit

- ▶ Ethisch-rechtl. Kontrolle von KI-Systemen: wichtiges Zukunftsthema
- ▶ Adressiert Herausforderungen auch in Richtung starke KI
- ▶ Grundlagen- und Anwendungsorientierte Forschung **jetzt fördern!**
- ▶ Erforderliche **interdisziplinäre Kompetenz jetzt aufbauen!**
- ▶ Sehr **kritische KI-Anwendungen vorerst entschleunigen!**

Ethics

Publikationen (siehe auch <http://christoph-benzmueller.de> → Publications)

- ▶ **Designing Normative Theories of Ethical Reasoning: Formal Framework, Methodology, and Tool Support** (Christoph Benz Müller, Xavier Parent, Leendert van der Torre), In Submitted, pp. 1–51, 2019. (Preprint: <https://arxiv.org/abs/1903.10187>)
- ▶ **Computer-supported Analysis of Arguments in Climate Engineering** (David Fuenmayor, Christoph Benz Müller), In CLAR 2020 – 3rd International Conference on Logic and Argumentation, Springer Nature Switzerland AG, Logic in Asia: Studia Logica Library, 2020. (Preprint <https://www.researchgate.net/publication/338829452>)
- ▶ **Harnessing Higher-Order (Meta-)Logic to Represent and Reason with Complex Ethical Theories** (David Fuenmayor, Christoph Benz Müller), In PRICAI 2019: Trends in Artificial Intelligence, Springer International Publishing, Lecture Notes in Artificial Intelligence, 2019. (Preprint <http://arxiv.org/abs/1903.09818>)
- ▶ **Universal (Meta-)Logical Reasoning: Recent Successes** (Christoph Benz Müller), In Science of Computer Programming, volume 172, pp. 48-62, 2019. (Preprint: <http://doi.org/10.13140/RG.2.2.11039.61609/2>)
- ▶ **A Deontic Logic Reasoning Infrastructure** (Christoph Benz Müller, Xavier Parent, Leendert van der Torre), In 14th Conference on Computability in Europe, CiE 2018, Kiel, Germany, July 30-August, 2018, Proceedings, Springer, LNCS, volume 10936, pp. 60-69, 2018. (Preprint: <https://tinyurl.com/y7tgoft9>)