

Ethisch-rechtliche Kontrolle autonomer Systeme – Machbar?

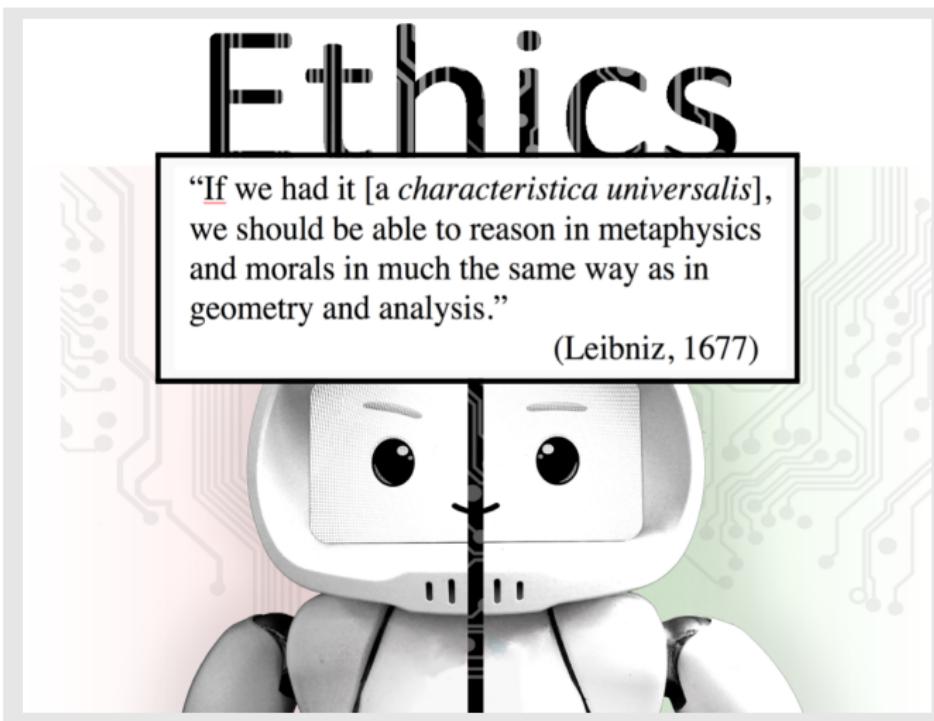
Prof. Dr. Christoph Benzmüller

FU Berlin | U Luxemburg

Ethics

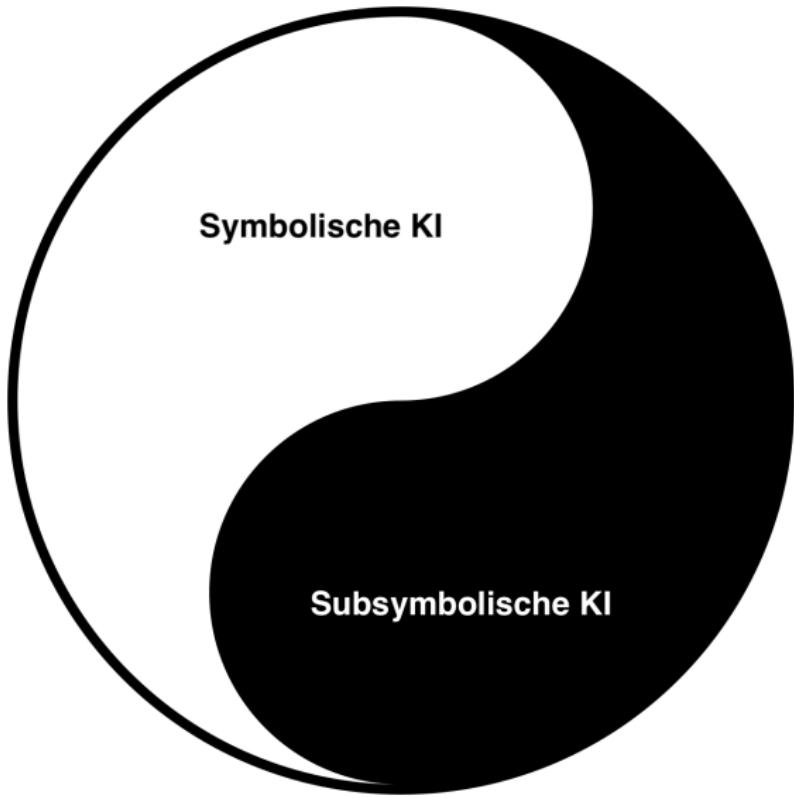
“If we had it [a *characteristica universalis*], we should be able to reason in metaphysics and morals in much the same way as in geometry and analysis.”

(Leibniz, 1677)

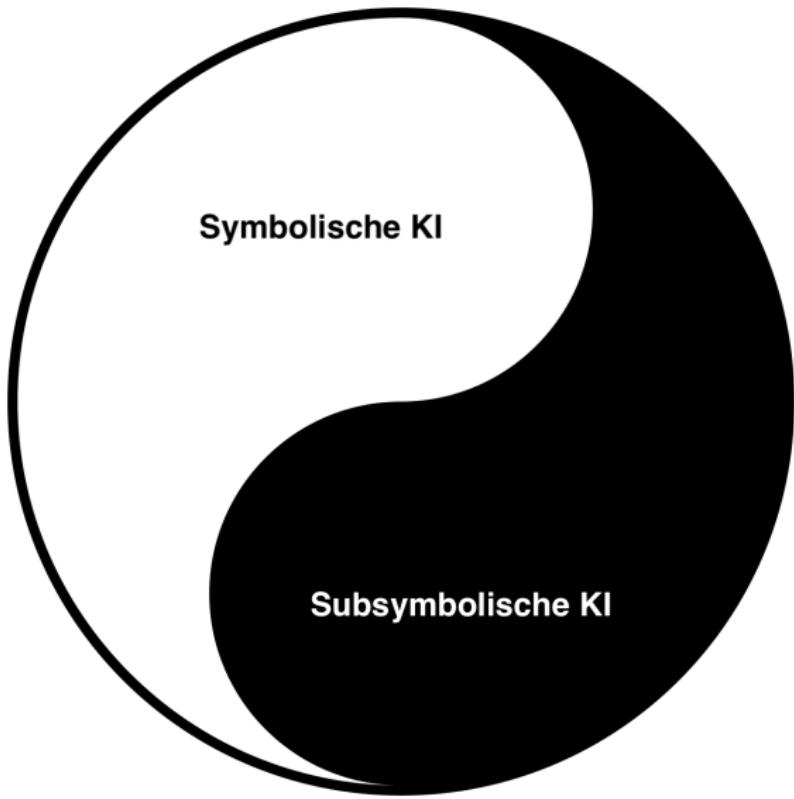


Artificial Intelligence in Automotive, München, 2.-4. Nov 2019

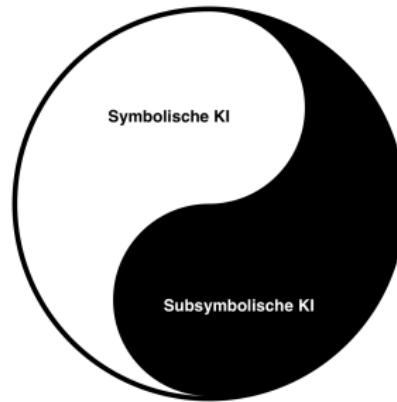
Yin und Yang der KI



Yin und Yang der KI — Erfolge & Medienpräsenz



Yin und Yang der KI — Erfolge & Medienpräsenz

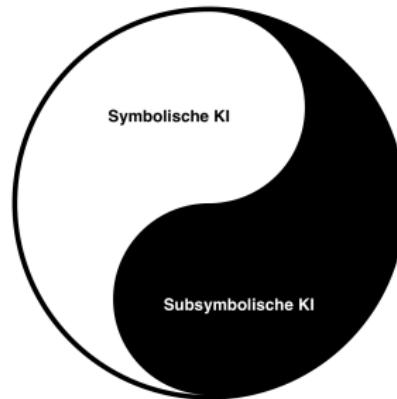


Erfolge

- ▶ 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**)

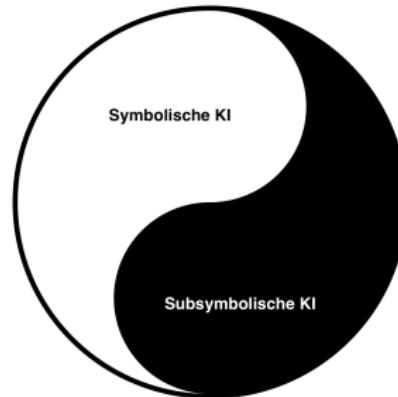


Yin und Yang der KI — Erfolge & Medienpräsenz

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)

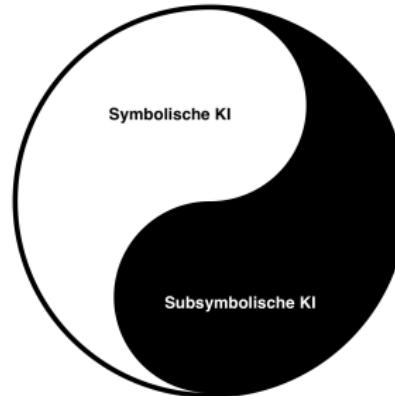


Yin und Yang der KI — Erfolge & Medienpräsenz

Erfolge

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

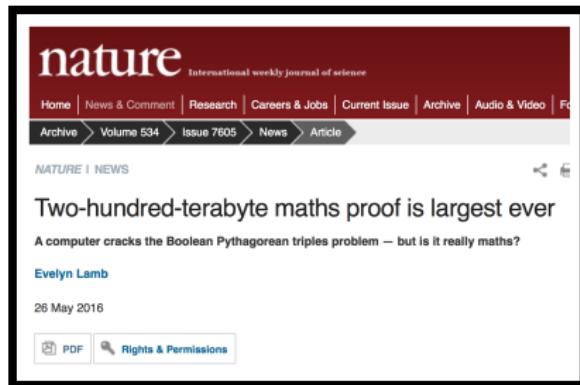
(symbolische KI)



Erfolge

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

(subsymbolische KI)



The screenshot shows the homepage of the journal 'nature'. The top navigation bar includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and a search bar. Below the navigation, a breadcrumb trail shows the path: Archive > Volume 534 > Issue 7605 > News > Article. The main headline reads: 'Two-hundred-terabyte maths proof is largest ever'. A sub-headline below it says: 'A computer cracks the Boolean Pythagorean triples problem — but is it really maths?'. The author's name, Evelyn Lamb, is listed, along with the publication date, 26 May 2016. At the bottom, there are links for PDF and Rights & Permissions.

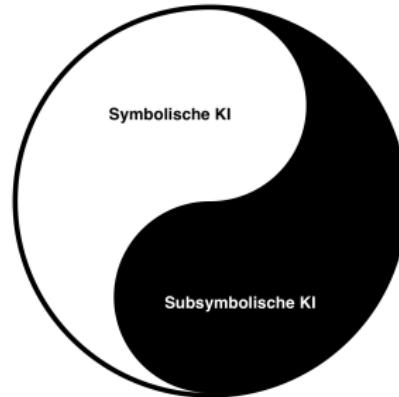


Yin und Yang der KI

Erfolge

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

(symbolische KI)



Erfolge

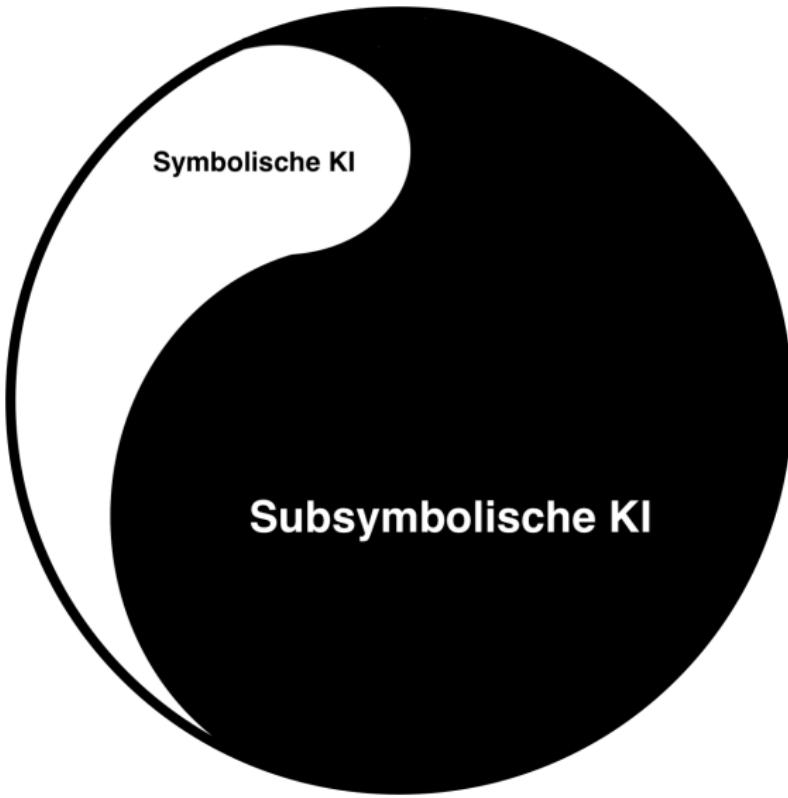
- ▶ Spiele:
Go und Schach
- ▶ endliche
Problemdomäne
- ▶ Weltmeisterniveau
erreicht: AlphaGo
und AlphaZero

(subsymbolische KI)

The screenshot shows a news article from SPIEGEL ONLINE. The headline reads "Der längste Mathe-Beweis der Welt". Below the headline, it says "Drei Mathematiker haben ein Zahlenrätsel geknackt - mithilfe eines Supercomputers. Der Beweis umfasst 200 Terabyte. Sie wollen wissen, worum es geht? Okay, versuchen wir es." The author's name, Holger Dambeck, is mentioned at the bottom.



Yin und Yang der KI — Ungesunder Hype?!



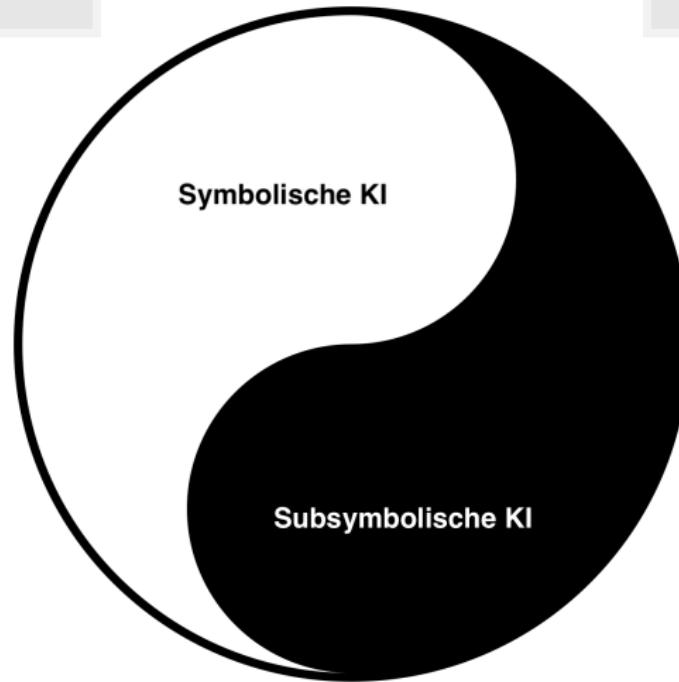
Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Sprache

...

Korrelationen
Muster
Robustheit
Lernen

...



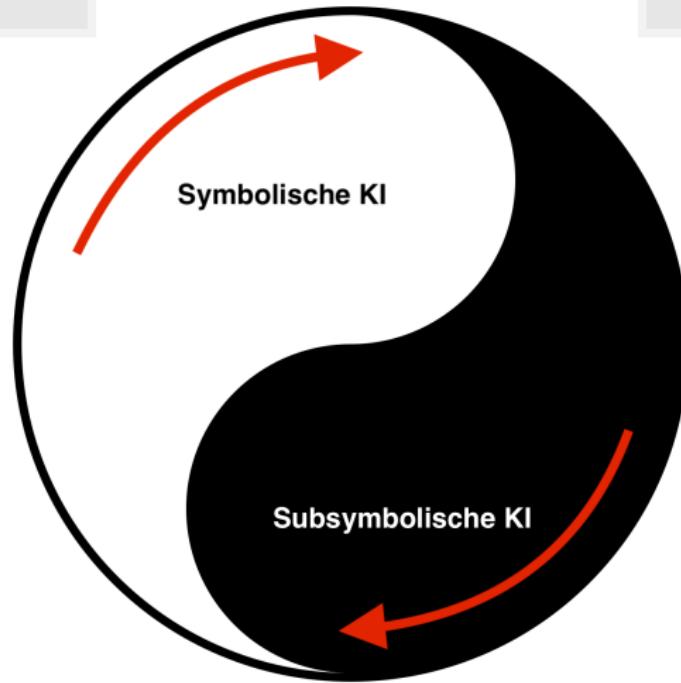
Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Sprache

...

Korrelationen
Muster
Robustheit
Lernen

...



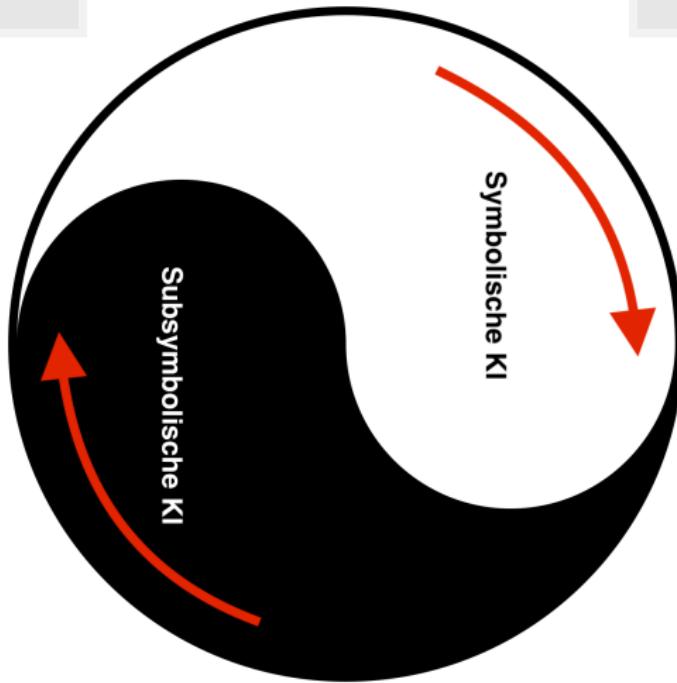
Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Sprache

...

Korrelationen
Muster
Robustheit
Lernen

...



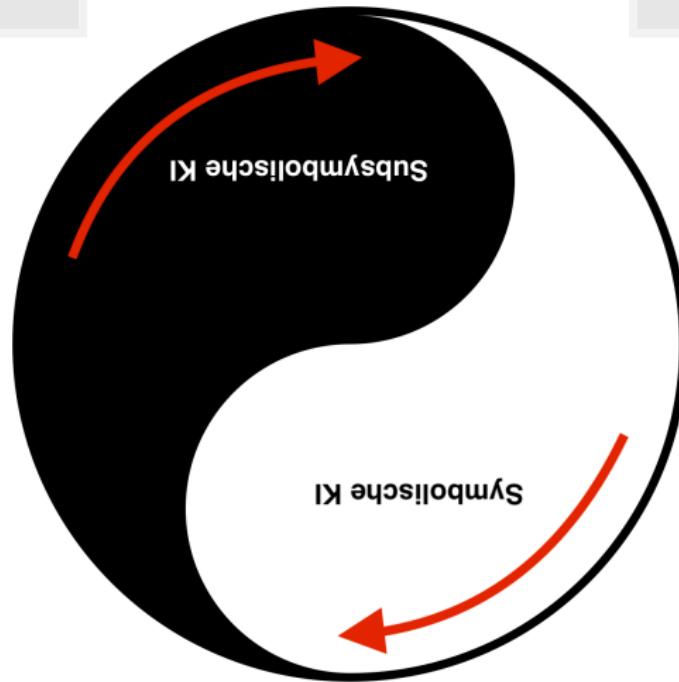
Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Sprache

...

Korrelationen
Muster
Robustheit
Lernen

...



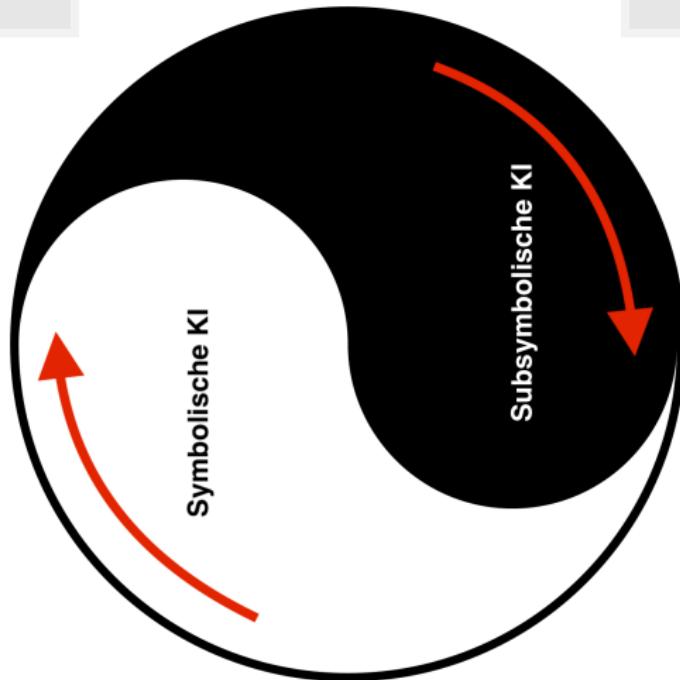
Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Sprache

...

Korrelationen
Muster
Robustheit
Lernen

...



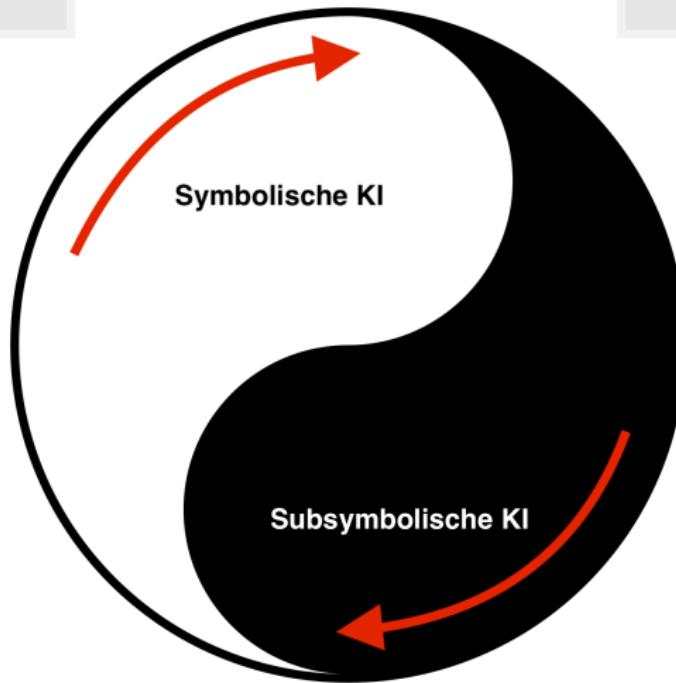
Yin und Yang der KI — The Next (really) Big Thing?!

Präzises Schließen
Abstraktion
Kausalität
Sprache

...

Korrelationen
Muster
Robustheit
Lernen

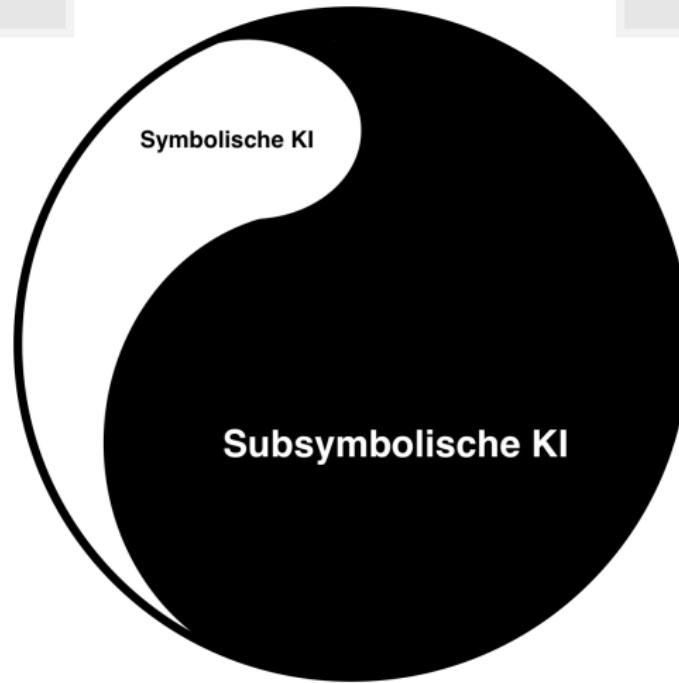
...



Datengetriebene ‘Beweise’ in der Mathematik ???

Präzises Schließen
Abstraktion
Kausalität
Sprache
...

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!

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!

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

$A \cup B := \dots$

$A \cap B := \dots$

$A \subseteq B \Leftrightarrow \dots$

$A = B := \dots$

\dots

\dots

Logik-Regeln

Annahmen

Beweisschritte

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

Theorem

Deduktiver Beweis — Mathematik

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

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

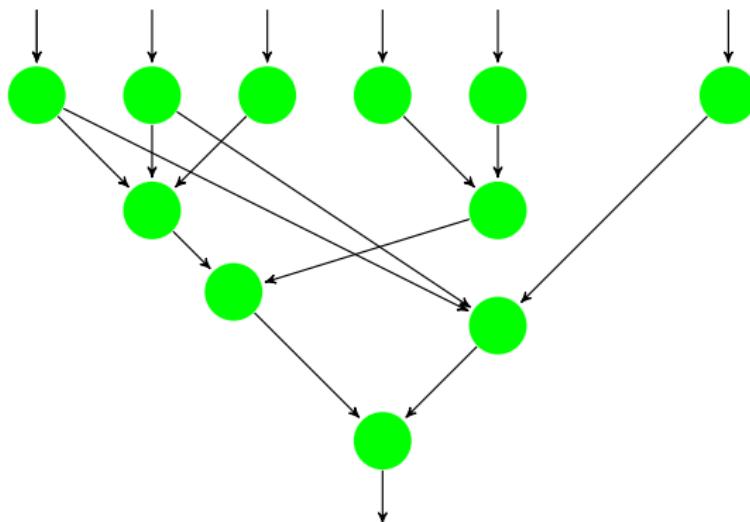
$$A \subseteq B \Leftrightarrow \dots$$

...

...

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

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

$A \cup B := \dots$
 $A \cap B := \dots$
 $A \subseteq B \Leftrightarrow \dots$
 $A = B \Leftrightarrow \dots$
...
...
Logik-Regeln

Annahmen

Theorembeweiser:

Computerprogramme
... die solche

Unser Beweiser Leo-
... für (polymathe)

Weitere Anwendungen
... Software/Hardware
... Rationale A

[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](#) | [Zeilgeist](#) | [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.

'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

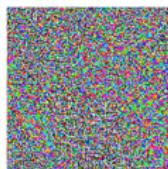


Adversarial Attacks:



“panda”

$+ .007 \times$



noise

=



“gibbon”

57.7% confidence

99.3% confidence

Example of an adversarial perturbation (Goodfellow et al., 2015)

'Beweise' in der Künstlichen Intelligenz

Was wird eigentlich gelernt?



person

0.88



person

0.90

Predictions from convolutional DNN on images of Kim Kardashian (Bourdakos, 2017)

'Beweise' in der Künstlichen Intelligenz



'Beweise' in der Künstlichen Intelligenz



Wenn es schief geht brauchen wir:

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

‘Beweise’ in der Künstlichen Intelligenz



‘Beweise’ in der Künstlichen Intelligenz

Superintelligenz:

Superintelligenz noch nicht in Sicht

leicht überzogene Erwartungen an datengetriebene KI

sehr hohe Dynamik im Gebiet



‘Beweise’ in der Künstlichen Intelligenz

Superintelligenz:

Superintelligenz noch nicht in Sicht

leicht überzogene Erwartungen an datengetriebene KI

sehr hohe Dynamik im Gebiet

Gerade deshalb sind die Herausforderungen groß:

Trustworthy AI made in Germany/Europe

‘Beweise’ in der Künstlichen Intelligenz

Superintelligenz:

Superintelligenz noch nicht in Sicht

leicht überzogene Erwartungen an datengetriebene KI

sehr hohe Dynamik im Gebiet

Gerade deshalb sind die Herausforderungen groß:

Trustworthy AI made in Germany/Europe

Was wir brauchen (in kritischen Anwendungen):

deduktive Kontrolle von ethischen-rechtlichen Vorgaben

weil: “Erlernen” solcher Vorgaben ist problematisch

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!**—
- ▶ Lösung durch "human in the loop"? —**Wohl kaum!**—

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!**—
- ▶ Lösung durch “*human in the loop*”? —**Wohl kaum!**—

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!**—
- ▶ Lösung durch “*human in the loop*”? —**Wohl kaum!**—

Moral Machine Experiment (siehe Nature, vol. 563)

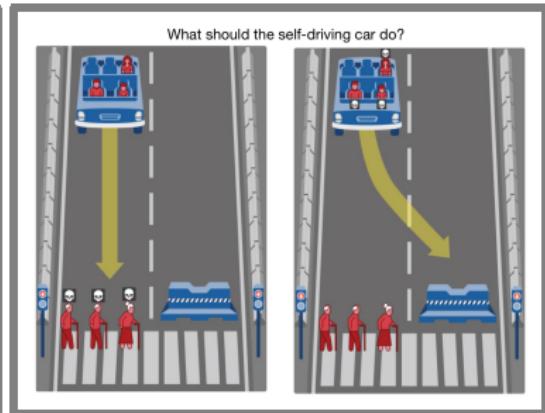
ARTICLE

<https://doi.org/10.1038/s41586-018-0637-6>

The Moral Machine experiment

Edmond Awad¹, Sohan Dsouza¹, Richard Kim¹, Jonathan Schulz², Joseph Henrich³, Azim Shariff^{3*}, Jean-François Bonnefon^{4*} & Iyad Rahwan^{1,5*}

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 — Einige Beispiele:

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

Gender: Frauen vor Männern (bei beiden Geschlechtern)

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

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

Teilweise im Widerspruch zu den Empfehlungen in:

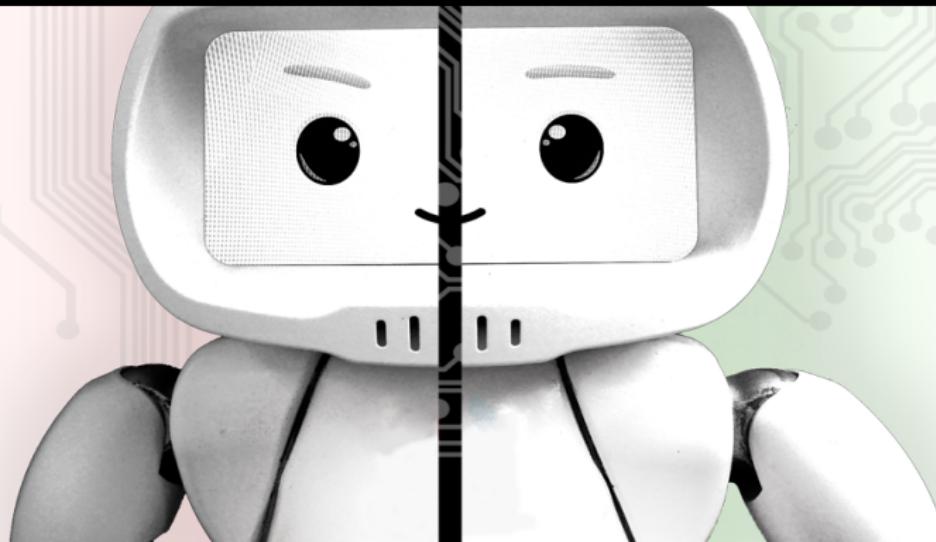
C. Luetge, **The German Ethics Code for automated and connected driving.**

Philos. Technol. 30, 547–558 (2017).

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**

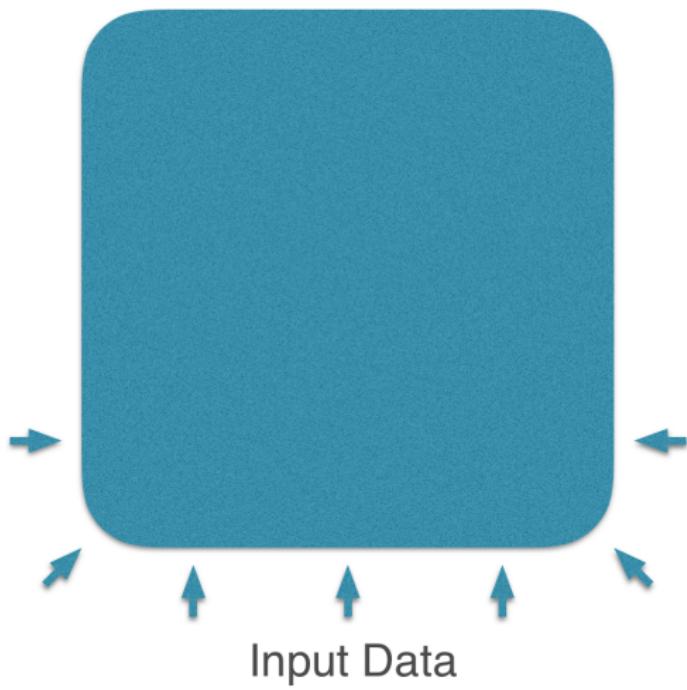
Existierende gesellschaftliche Prozesse basieren auf:

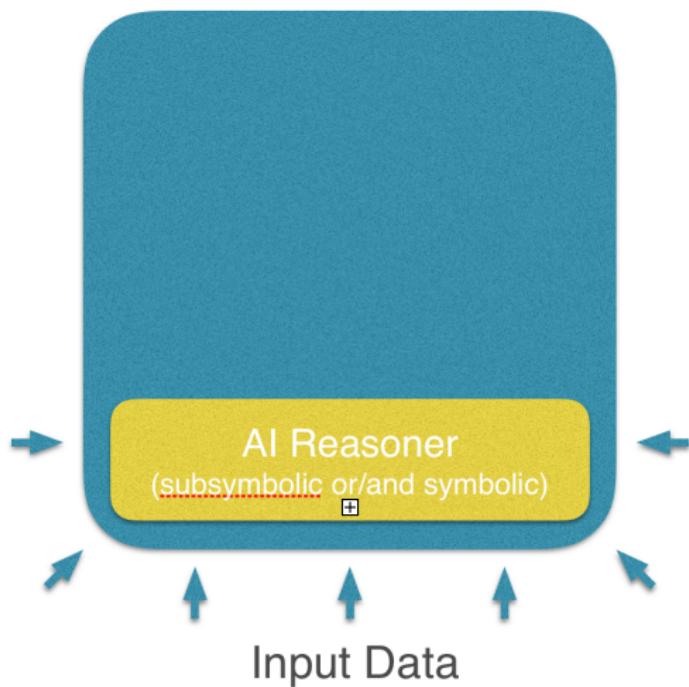
- ▶ **Erklärungen, rationaler Argumentation & Dialog,**
- ▶ inklusive **explizitem normativem Schließen** (rechtlich & ethisch)

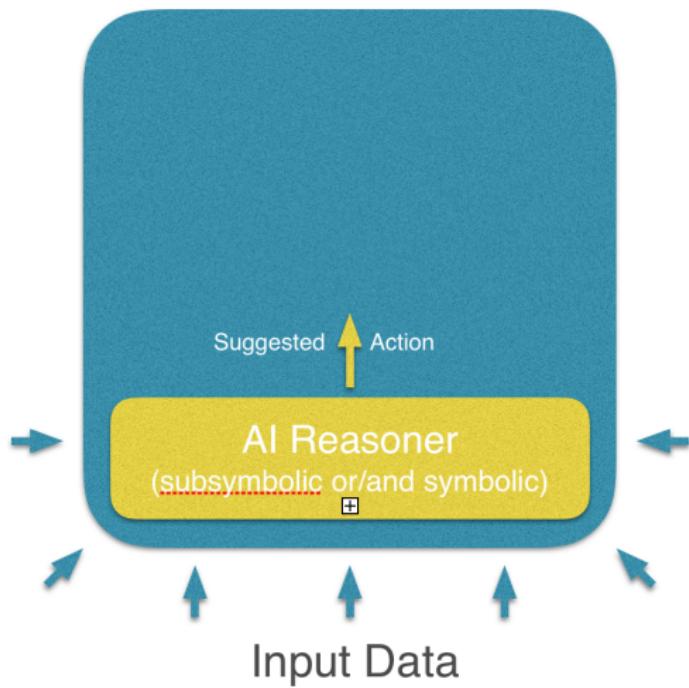
Entwicklung von IASs ohne solche Kompetenzen? Wie sinnvoll?

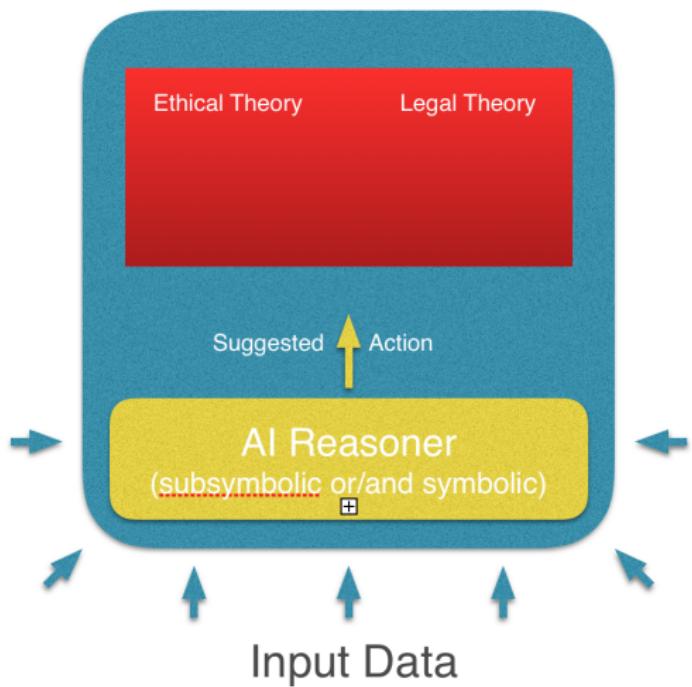


IAS



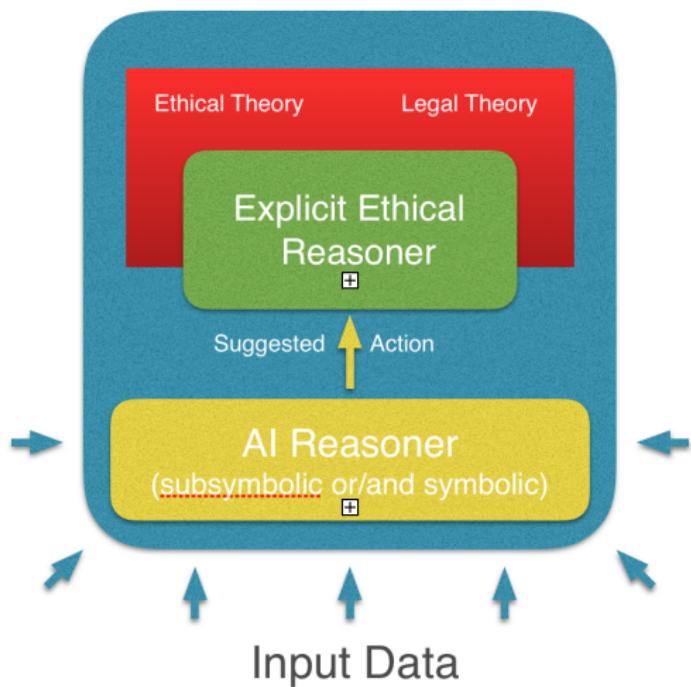






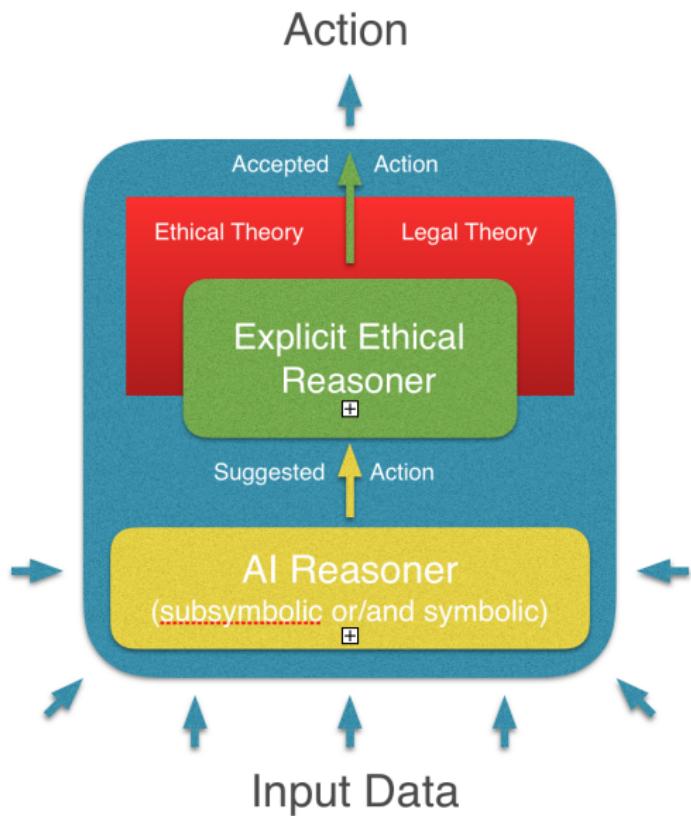
Pseudo-Ethischer KI Agent

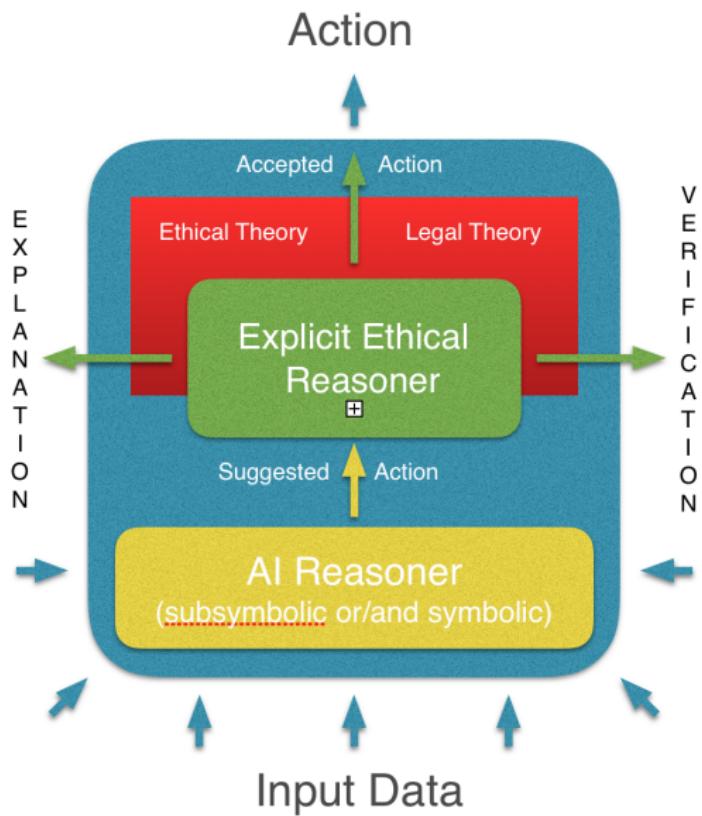
Trustworthy AI



Pseudo-Ethischer KI Agent

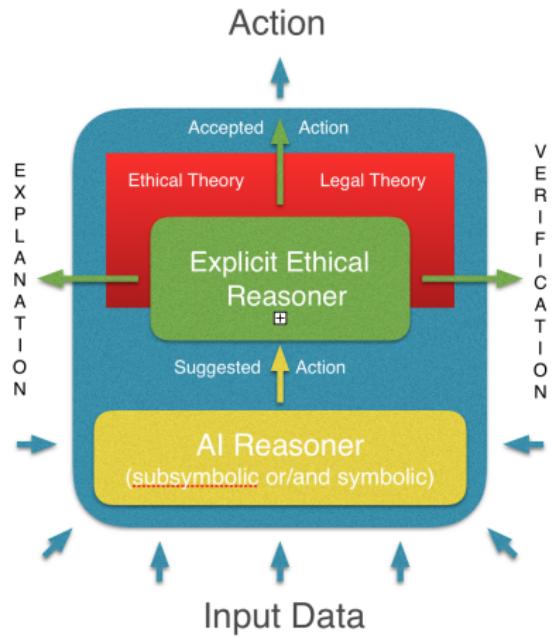
Trustworthy AI





Pseudo-Ethischer KI Agent

Trustworthy AI



Verwandte Arbeiten

- ▶ 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]

Bedeutung

Bundesregierung, Nov 2018: Strategie Künstliche Intelligenz

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

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

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

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

<https://tinyurl.com/y8h94ouv>

Bedeutung

Bundesregierung, Nov 2018: Strategie Künstliche Intelligenz

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

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

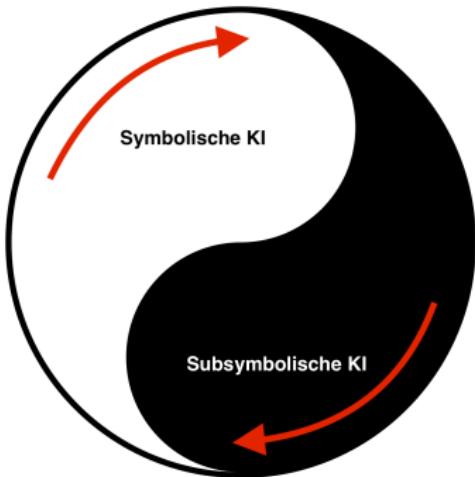
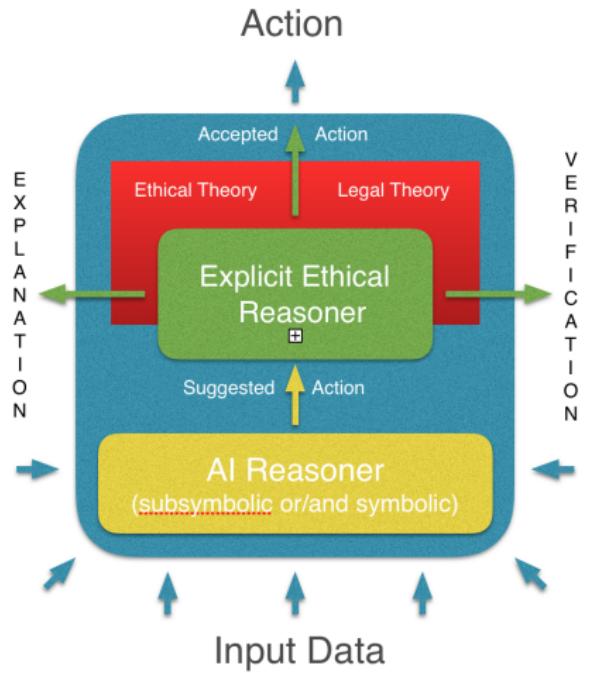
Ben Goertzel, CEO SingularityNET (zuvor Hanson Robotics); Nov 2018: "Toward Democratic, Lawful Citizenship for AIs, Robots, and Corporations"

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

<https://tinyurl.com/y8h94ouv>

Pseudo-Ethischer KI Agent

Trustworthy AI



Normatives Schließen

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

- ▶ Dilemmas, Paradoxien, inkompatible Theorien, etc.
- ▶ Geeignete Modellierung des Begriffs der **Obligation**
- ▶ **Obligation:** nicht-trivialer \Box -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: Welches Recht? Welche Ethik? **Welche Logik(en)?**

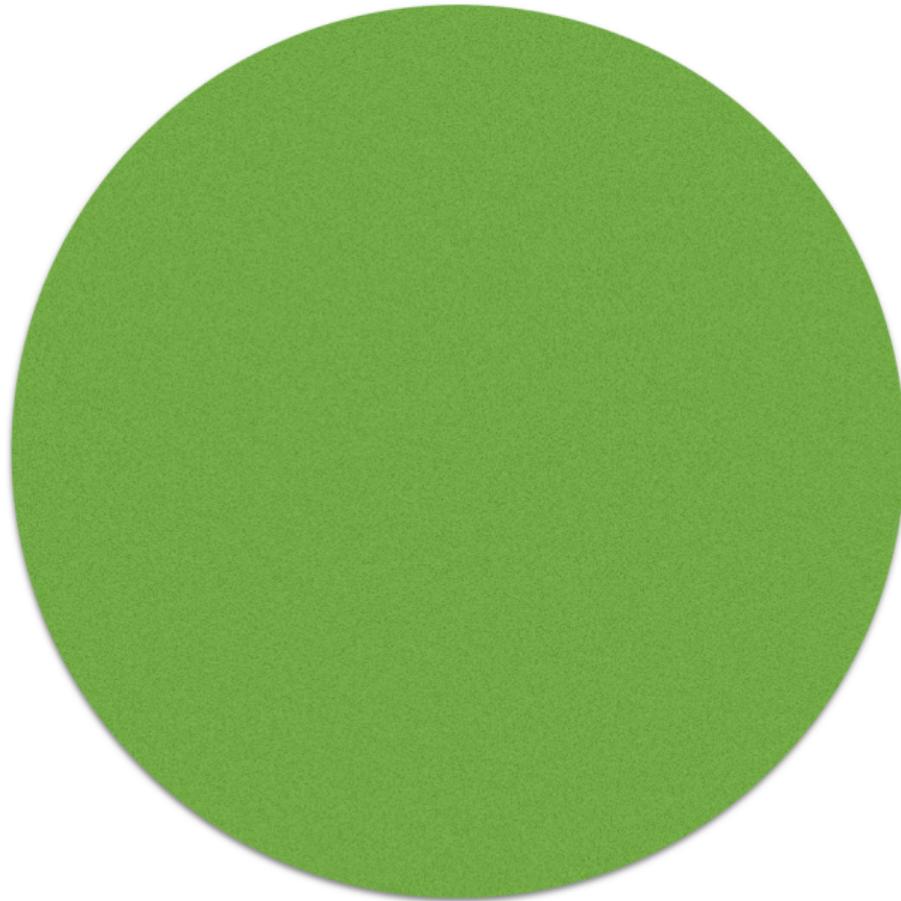
- ▶ Dilemmas, Paradoxien, inkompatible Theorien, etc.
- ▶ Geeignete Modellierung des Begriffs der **Obligation**
- ▶ **Obligation:** nicht-trivialer \Box -Operator der Modallogik/Deontischen Logik
- ▶ Problem: “Contrary-to-duty” (**CTD**) Szenarien

CTD Beispiel (X. Parent): EU GDPR

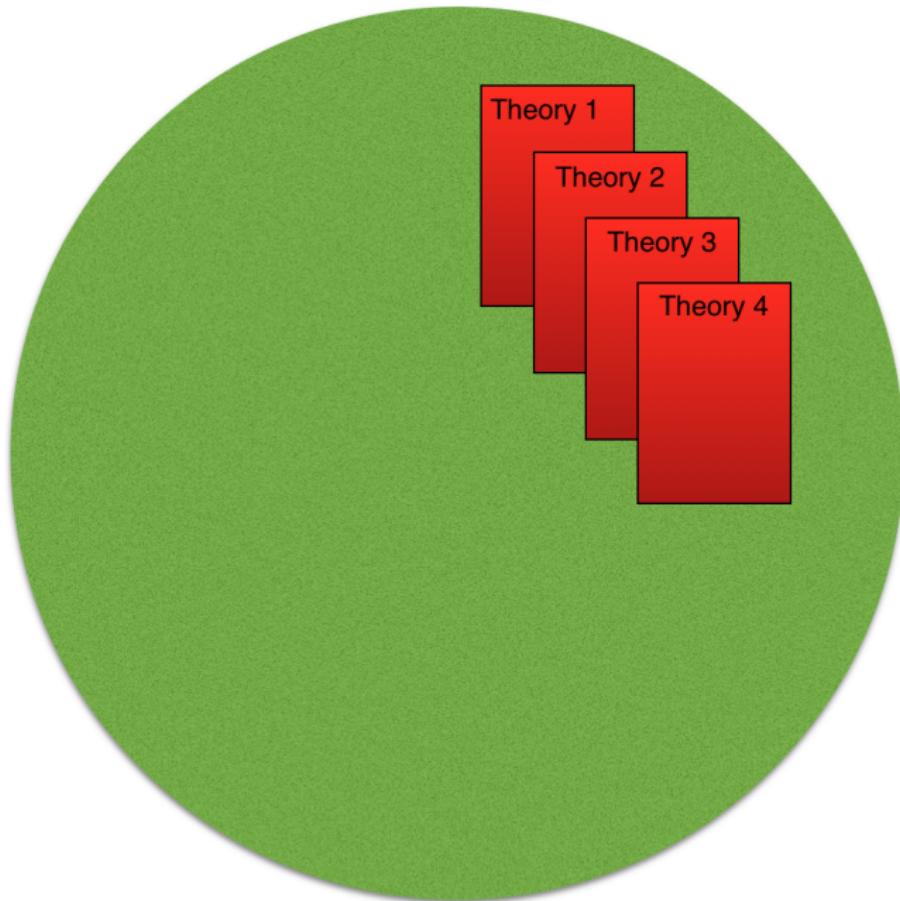
1. Personal data shall be processed lawfully. (Art. 5)
E.g., the data subject must have given consent to the processing. (Art. 6/1.a)
2. **Implizit:** The data shall be kept, for the agreed purposes, if processed lawfully.
3. If personal data has been processed unlawfully, the controller has the obligation 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: Paradoxe/Inkonsistenz — Ex falso quodlibet!

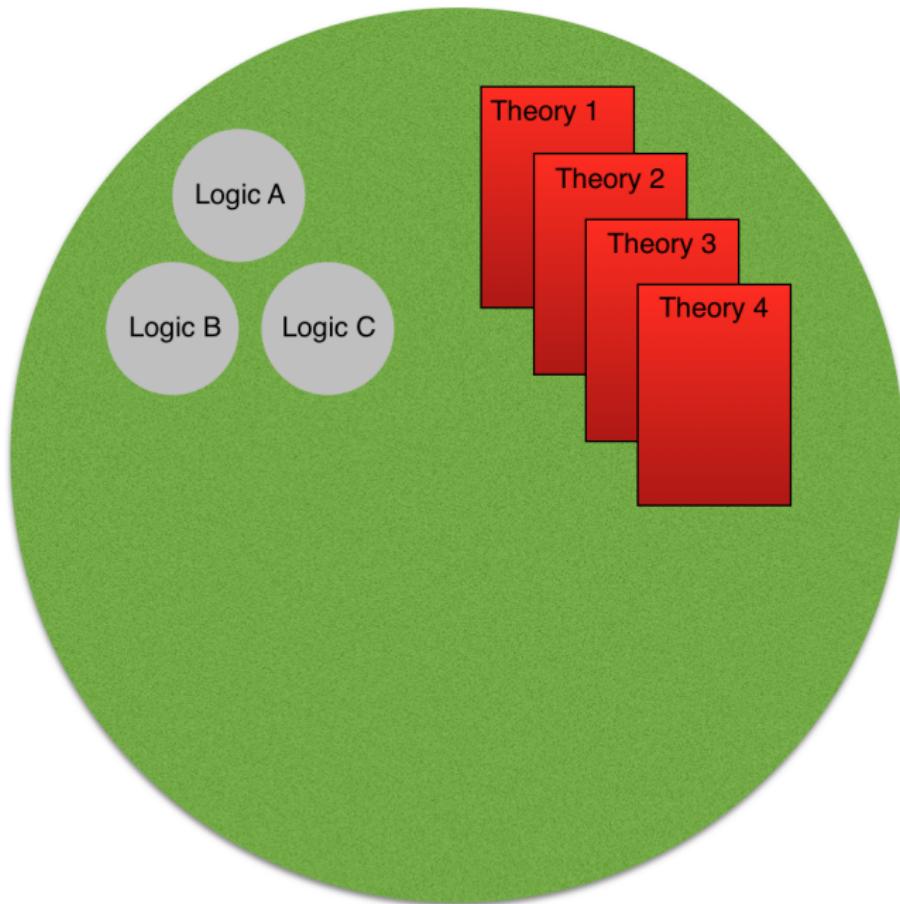
Experimentierplattform für Normatives Schließen und Maschinen-Ethik



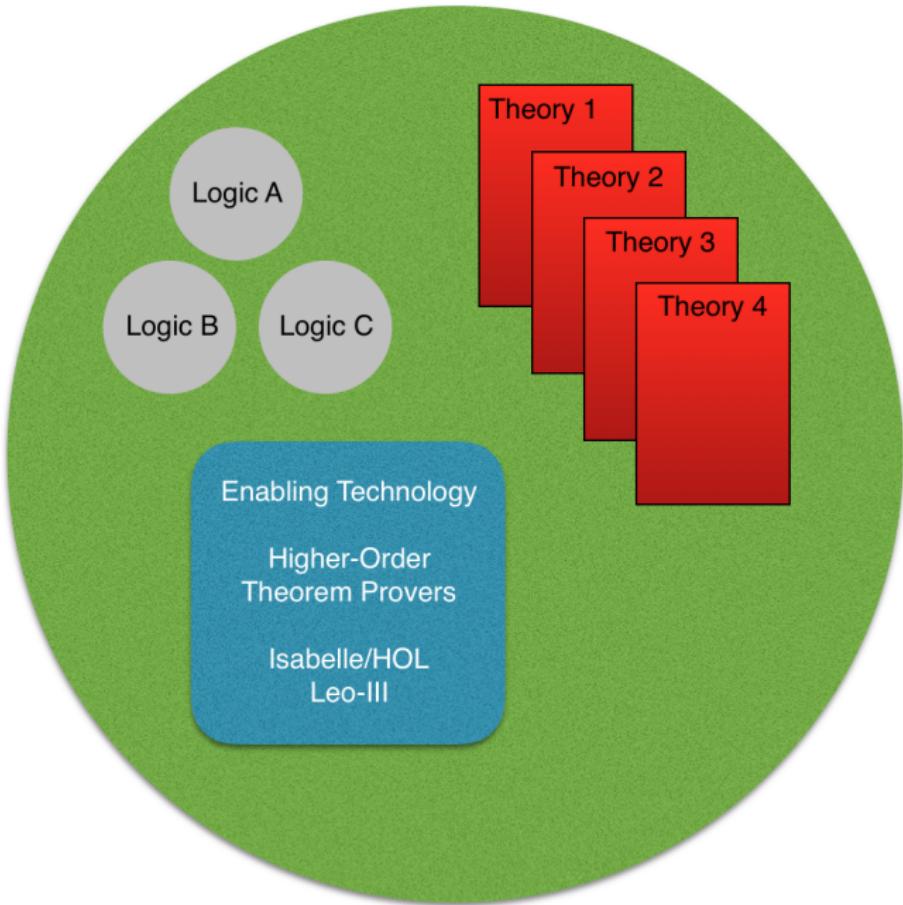
Experimentierplattform für Normatives Schließen und Maschinen-Ethik



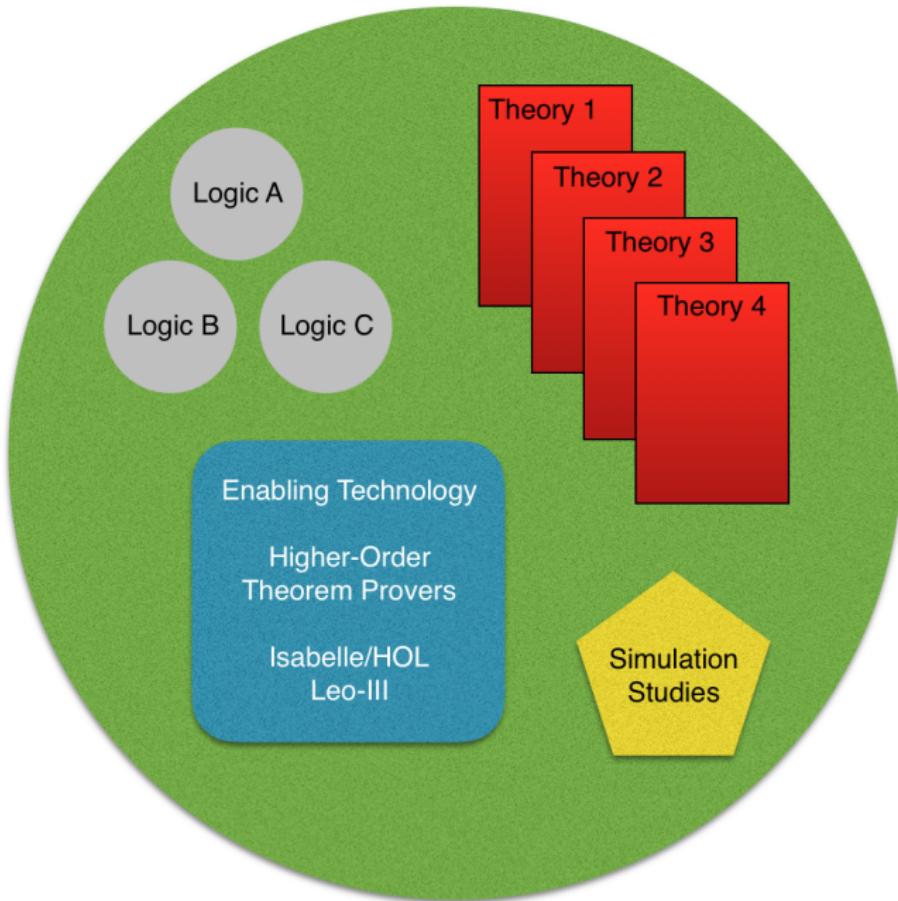
Experimentierplattform für Normatives Schließen und Maschinen-Ethik



Experimentierplattform für Normatives Schließen und Maschinen-Ethik



Experimentierplattform für Normatives Schließen und Maschinen-Ethik



Experimentierplattform für Normatives Schließen — Demo möglich!

The screenshot shows the Isabelle/HOL proof assistant interface. The main window displays a theory file named 'GDPR.thy'. The code defines a theory 'GDPR' that imports 'SDL'. It includes an axiomatization with obligations related to data processing lawfully, erase data, and kill a boss. Below this, there are sections for experiments using 'nitpick' and 'sledgehammer' to check consistency and derive Falsum. The bottom part of the interface shows the results of a 'sledgehammer' run, indicating that a proof was found. A vertical sidebar on the right provides navigation links for Documentation, Sidekick, State, and Theories.

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

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

Output Query Sledgehammer Symbols

Isabelle/HOL (Beweisassistent) als Werkzeug für Universelles Schließen

Experimentierplattform für Normatives Schließen — Demo möglich!

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

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

Gefahren-Zone:
Paradoxien and Inkonsistenzen!

Proof state Auto update Update Search: 100%

Parallele Entwicklung und Verifikation von
ethisch-rechtlichen Theorien ↔ passende Logikformalismen

Output Query Sledgehammer Symbols

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 Benzmüller

[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

 Springer Link

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

Home

Authors

Authors and affiliations

David Fuenmayor , Christoph Benzmüller

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

"Act in accordance with the principles of justice."
call this the Principle of Generic Consideration.
consideration of the features or good

BACHELOR'S THESIS

I shall
e formal
generic
y, 1978)

A
"Behandl
"Was du nic

Modelling the US Constitution in HOL

On establishing a dictatorship with Gödel

llst."
ern zu."



Valeria Zahoransky

FOR

L

supervised and examined by
Prof. Dr. Christoph BENZMÜLLER

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

Home

Ethics

Fazit

- ▶ KI Hype etwas einseitig überzogen; dennoch gilt . . .
 . . . KI-Technologie hat enormes, positives Potential.
- ▶ Vor den Problemen sollten wir nicht die Augen verschließen,
 eine “deploy-and-then-regulate”-Mentalität ist nicht angebracht.
- ▶ Entsprechende Studien zur Kontrolle von KI-Systemen jetzt starten!
- ▶ Erforderliche interdisziplinäre Kompetenz jetzt aufbauen!



Vorschlag: KI-Zukunftslabor



Projekttitle: ENoRME: Explicit Normative Reasoning and Machine Ethics

Beantragte Laufzeit: 36 Monate

Angestrebter Standort: Berlin

Forschungs- und Entwicklungsvorhaben:

“Cloud-basierte on-demand” Plattform zur Modellierung und Automatisierung ethisch-rechtlicher Theorien

- ▶ Methoden- und Technologieentwicklung
Theorie—Entwurf—Implementierung—Anwendungsstudien
- ▶ Lehrveranstaltungen, Workshops, Konferenzen, Tutorien und e-Learning
- ▶ Öffentlichkeitsarbeit und Community Management
- ▶ Kooperation mit nationalen und internationalen KI-Initiativen

Vorschlag: KI-Zukunftslabor



- ▶ Methoden- und Technologieentwicklung
Theorie—Entwurf—Implementierung—Anwendungsstudien
- ▶ Lehrveranstaltungen, Workshops, Konferenzen, Tutorien und e-Learning
- ▶ Öffentlichkeitsarbeit und Community Management

Vorschlag: KI-Zukunftslabor



- ▶ Methoden und Tools
- ▶ Lernmethoden und e-Learning
- ▶ Öffentlichkeitsarbeit und Community Management
- ▶ Kooperation mit nationalen und internationalen KI-Initiativen

Vorschlag: KI-Zukunftslabor



**Mehrwert zu laufenden Forschungs- und Entwicklungsarbeiten:
“Interdisziplinäre Forschung an Schnittstellen”:**

- (i) Automatisches Theorembeweisen und Deduktion
 - (ii) Universelles Logisches Schließen
 - (iii) Modellierung ethisch-rechtlicher Theorien und Normatives Schließen
 - (iv) Maschinen-Ethik
 - (v) Robotik
 - (vi) Maschinelles Lernen
- ▶ Bindeglied zwischen Forschung und Anwendung
 - ▶ Katalysator für praktische Entwicklungen
 - ▶ Internationales Netzwerk / Integration von Communities
 - ▶ Lehr-/Lernmaterial
 - ▶ Neue Generation von Wissenschaftlern

Vorschlag: KI-Zukunftslabor



Partizipierende Partnereinrichtungen und Wissenschaftler/innen:

FU Berlin: (1) C. Benzmüller, (2) Raúl Rojas, (3) Daniel Göhring

U Koblenz: (4) Claudia Schon (Uli Furbach)

U Luxembourg: (5) Xavier Parent, (6) Alexander Steen (Leon van der Torre)

U Bergen: (7) Marija Slavkovik

U Miami: (8) Geoff Sutcliffe

U Liverpool: (9) Louise Dennis (Michael Fisher)

U Campinas: (10) Walter Carnielli, (11) Juliana Bueno-Soler

U Buenos Aires: (12) Maria Vanina Martinez

Weitere akademische Partner: Toby Walsh (UNSW Sydney), Beishui Liao (Zhejiang U), Adam Pease (Infosys, Palo Alto), Jan Broersen (U Utrecht), Stephan Schulz (DHBW Stuttgart), Ralf Romeike (FU Berlin)

Industrie & Forschungseinrichtungen: Latentine GmbH, wizAI solutions GmbH, LuxAI S.A., KPMG Lighthouse Luxembourg, DLR;

Zusammenarbeit: Research Center MATH+; JFK-Institute

Vorschlag: KI-Zukunftslabor



2020												2021												2022											
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Christoph Benzmüller, FU Berlin, DE, W3																																			
Raúl Rojas, FU Berlin, DE, W3																																			
Daniel Göring, FU Berlin, DE, W2																																			
Claudia Schon, U Koblenz-Landau, DE, W2																																			
Xavier Parent, U Luxembourg, LU, W2																																			
Alexander Steen, U Luxembourg, LU, W1																																			
Marija Slavkovik, U Bergen, NO, W2																																			
Geoff Sutcliffe, U Miami, US, W3																																			
Louise Dennis, U Liverpool, UK, W2																																			
Walter Carnielli, U Campinas, BR, W3																																			
Juliana Bueno-Soler, U Campinas, BR, W2																																			
Maria Vanina Martinez, U Buenos Aires, AR, W2																																			
David Fuenmayor, Implementation of the Cloud-Based URW, Young Talented Researcher, DE, E13																																			
Explanations:												DE	EU	other	9 Month Core Period												female	male							

Vorschlag: KI-Zukunftslabor



Besondere Eignung der Wissenschaftler/innen für ENoRME:

- (1) Universelles Schließen, Deduktionssysteme, Maschinen-Ethik
- (2) KI, Robotik, Maschinelles Lernen
- (3) Autonomes Fahren, Robotik, Maschinelles Lernen
- (4) Beschreibungslogiken, Theorembeweisen, Maschinen-Ethik, Ontologien
- (5) Normatives Schließen, Deontische Logiken, Ethisch-Rechtliche Theorien
- (6) Automatisches Theorembeweisen, Systementwicklung
- (7) Kollektives Schließen, Maschinen-Ethik
- (8) Infrastrukturen und Testumgebungen für Deduktionssysteme
- (9) Autonome Systeme, Verifizierbare Systeme, Maschinen-Ethik
- (10) Nichtklassische Logiken, Logikkombinationen, Parakonsistentes Schließen
- (11) Probabilistisches Parakonsistentes Schließen
- (12) Schließen bei Unsicherheit & Inkonsistenz, Semantic Web

Vorschlag: KI-Zukunftslabor



Projektziele und erwartete Resultate:

- Cloud-basierte Werkbank und Infrastruktur
- Standardisierung und Benchmarking im Normativen Schließen
- Theorie und Implementierung neuer, relevanter Logikkombinationen
- Beispiele automatisierter ethisch-rechtlicher Theorien
- Spezielle Technologien (Beweisen und Modelle) im Normativen Schließen
- Agenten-basierte Simulationsplattform
- Fallstudien
- Entwurf und Modellierung von Architekturen für “Ethical Governors”
- Exemplarische Implementierungen:
 - (i) Autonome Fahrzeuge, (ii) Finanz- und Bankwesen, (iii) Soziale Robotik,
 - (vi) Pharmazie and Gesundheitswesen, (v) Kognitive Robotik, (vi) Marketing
- Konferenzen, Workshops, und Tutorien
- Vorlesungen, Seminare, e-Learning, Projekte
- Öffentlichkeitsarbeit

Vorschlag: KI-Zukunftslabor



Projektziele und erwartete Resultate:

- Cloud-basierte Werkbank und Infrastruktur
- Standardisierung und Benchmarking im Normativen Schließen
- Theorie und Implementierung neuer, relevanter Logikkombinationen
- Beispiele automatisierter ethisch-rechtlicher Theorien
- Spezielle Technologien (Beweisen und Modelle) im Normativen Schließen
- Agentenarchitekturen
- Falls ... unfortunately funded by BMBF
- Entwurf und Modellierung von Architekturen für "Ethical Governors"
- Exemplarische Implementierungen:
 - (i) Autonome Fahrzeuge, (ii) Finanz- und Bankwesen, (iii) Soziale Robotik,
 - (vi) Pharmazie and Gesundheitswesen, (v) Kognitive Robotik, (vi) Marketing
- Konferenzen, Workshops, und Tutorien
- Vorlesungen, Seminare, e-Learning, Projekte
- Öffentlichkeitsarbeit