

# Ethico-legal Governance of Intelligent Artificial Agents

Can post-hoc normative reasoning competencies prevent AI systems from going rogue?

Christoph Benzmüller

Freie Universität Berlin & Latentine GmbH

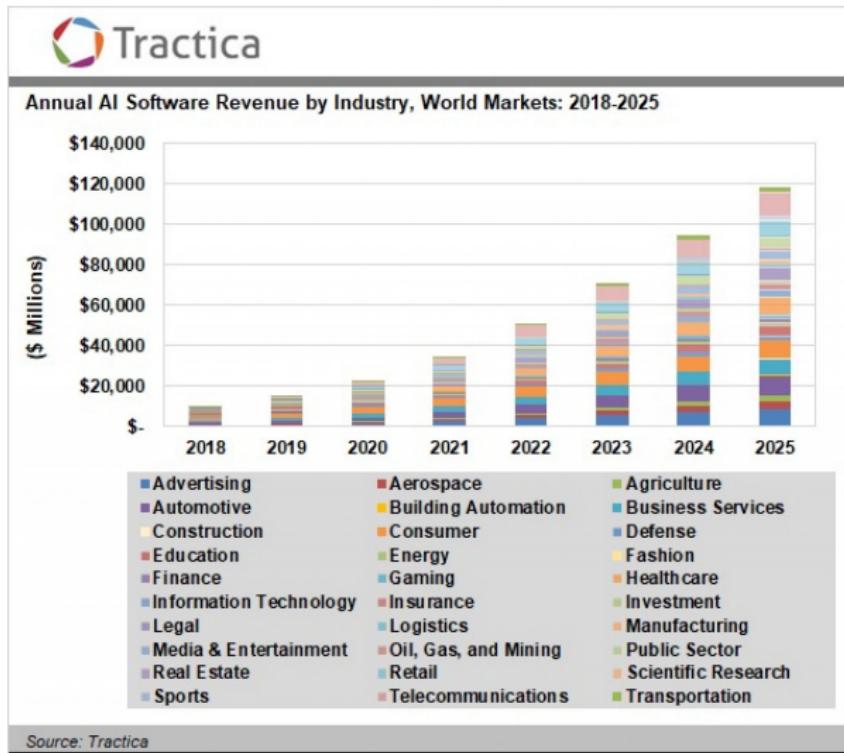


ZJULogAI, 26. Oktober, 2020

# Talk Structure

1. What is AI?
2. Own position
3. How to create trust & how to control?

# What is AI?



Steam engine of the 21st century?

# What is AI?

NEWS

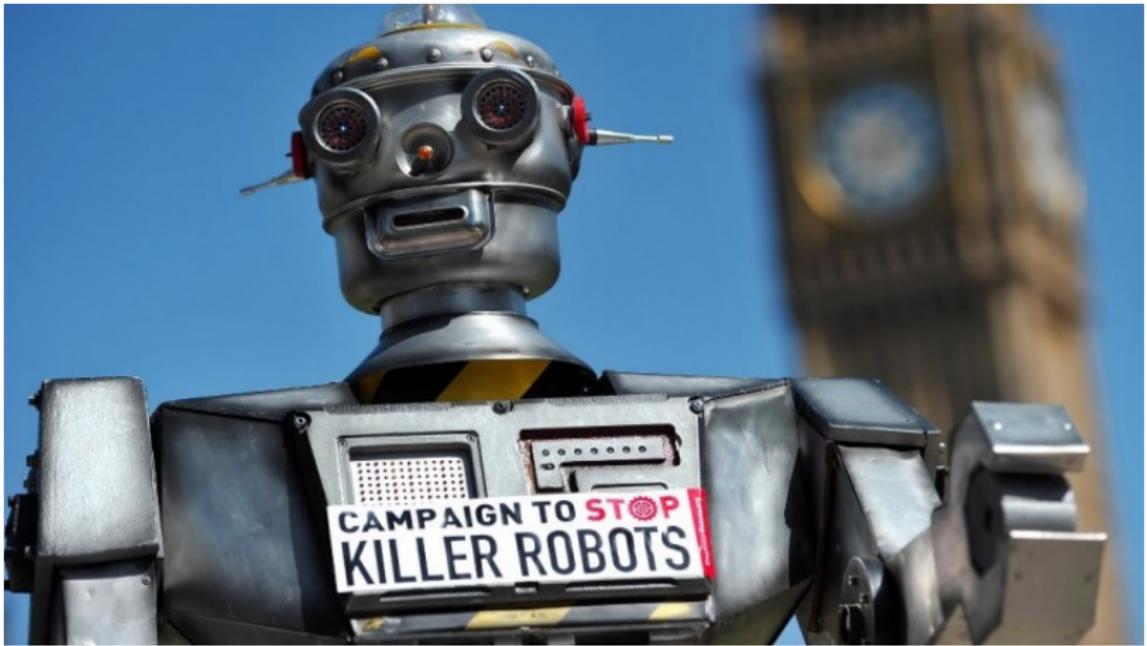
## Bitkom: Digitalization to wipe out millions of jobs in Germany

Millions of workers in Germany will likely lose their jobs and be replaced by robots and AI algorithms by 2023, German IT association Bitkom said in a study. The group urged politicians to take the issue more seriously.

Source: Deutsche Welle, Feb, 2018

A source for social tensions  
and diverging wealth distribution?

# What is AI?



Source: Deutschlandfunk/AFP, Carl Court, 2019

A danger regarding military escalation?

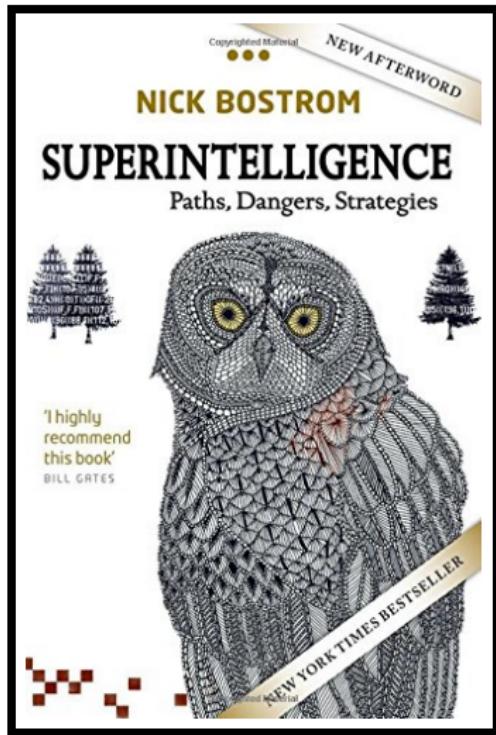
# What is AI?



Source/Photograph: Qiangjing Evening News

An interesting option for future partnerships?

# What is AI?



A new evolutionary step that leaves people behind?

# What is AI?

## How Innovative AI Solutions Can Help Combat Global Warming



**Asokan Ashok** Forbes Councils Member  
**Forbes Technology Council** COUNCIL POST | Paid Program  
Innovation

---

Source: Forbes, 2020

A tool for combatting our environmental sins?

# What is AI?

Artificial intelligence / Machine learning

## Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

by **Karen Hao**

June 6, 2019

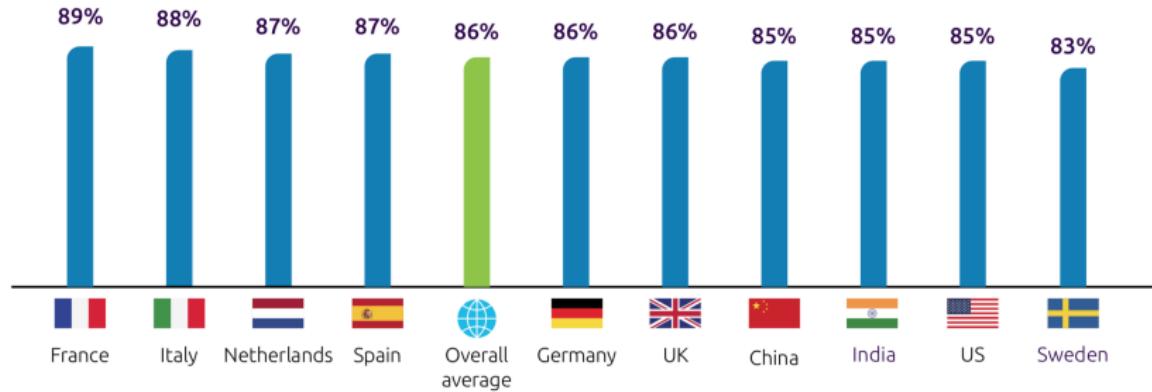
Source: MIT Technology Review, 2019

Itself the next big environmental sin?

# What is AI?

Figure 4. Nearly nine in ten organizations across countries have encountered ethical issues resulting from the use of AI

In the last 2-3 years, have the below issues resulting from the use and implementation of AI systems, been brought to your attention? (percentage of executives, by country)

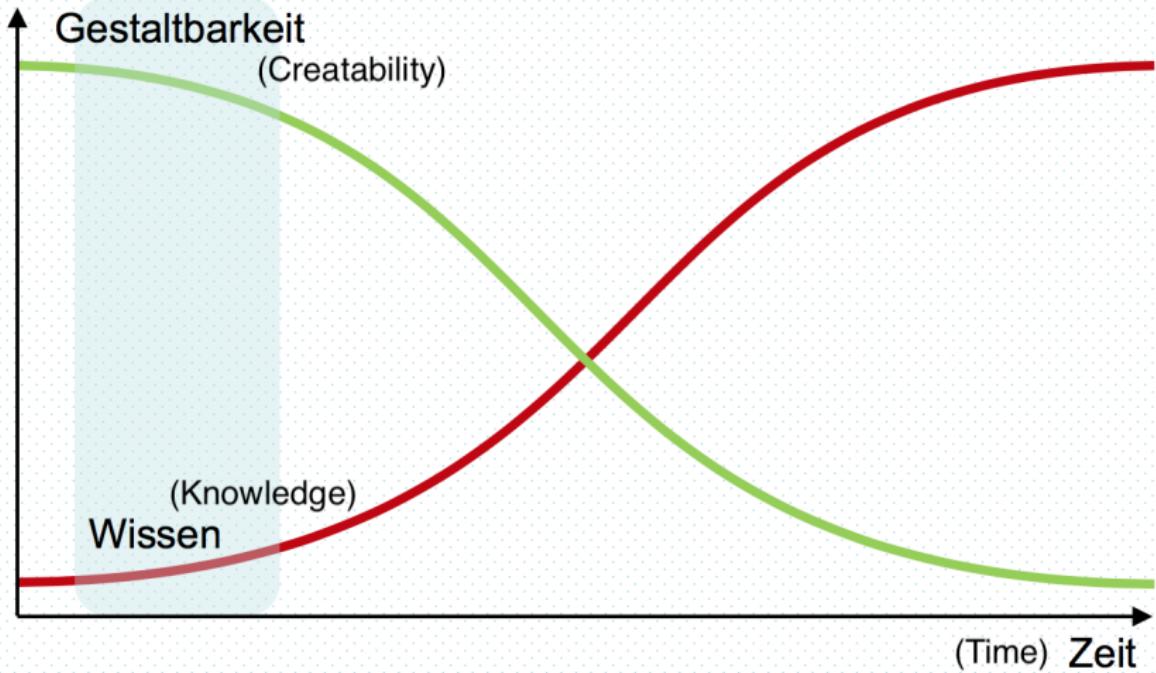


We presented over 40 cases where ethical issues could arise from the use of AI, to executives across sectors. We asked them whether they encountered these issues in the last 2-3 years.

Source: Capgemini Research Institute, Ethics in AI executive survey, N = 1,580 executives, 510 organizations.

## Source of ethical and legal conflicts?

# What is AI?



Challenge for active participation!

# What is AI? – Many definitions, no consensus

## Weak/Applied AI

Limited, 'smart' systems that simulate intelligence

vs.

## Strong AI

Machines with (at least) general human intelligence

## Cognitive Simulation

Exploring and testing theories of human cognition with computers

# What is AI? – Many definitions, no consensus

## Weak/Applied AI

Limited, 'smart' systems that simulate intelligence

vs.

## Strong AI

Machines with (at least) general human intelligence

### Cognitive Simulation

Exploring and testing theories of human cognition with computers

- ▶ *AI is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable. (McCarthy)*
- ▶ *The study of how to make computers do things at which, at the moment, people are better. (Rich and Knight)*
- ▶ ...
- ▶ *Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather it reflects a broader and deeper capability for comprehending our surroundings – 'catching on', 'making sense' of things, or 'figuring out' what to do. (Gottfredson, 1997)*

# What is AI? – Many definitions, no consensus

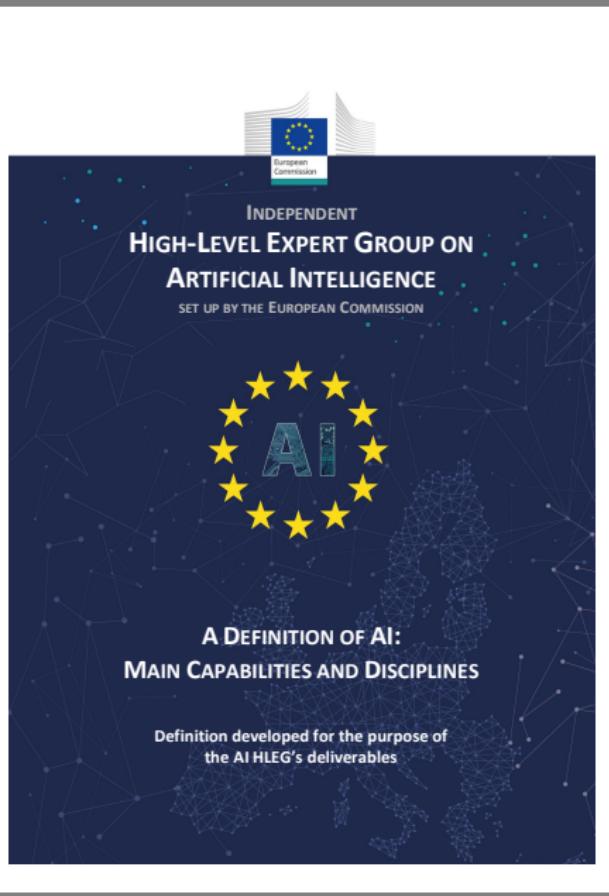
## Weak/Applied

Limited, 'smart' systems simulate intelligent behaviour

- ▶ *AI is the science and engineering of computer programs that can perform tasks that require human intelligence, such as solving complex problems or learning from experience.*
- ▶ *The study of how computers can be made to perform tasks that are normally done by people.* (Rich and Lewis, 2011)
- ▶ ...
- ▶ *Intelligence is a very general concept that includes the ability to reason, learn quickly and easily, apply academic skill, or have the capability for complex things, or 'figuring things out'.*

## Strong AI

Can (at least) general intelligence



s, especially intelligent computers to understand methods that are

moment, people are

er things, involves the understand complex ideas, learning, a narrow or and deeper 'making sense' of

# AI – Own working definition

## Def.: Artificial Intelligence

A science of computational technologies that are developed to achieve and explain **intelligent** behavior in machines.

# AI – Own working definition

## Def.: Artificial Intelligence

A science of computational technologies that are developed to achieve and explain **intelligent** behavior in machines.

## Def.: Intelligence

Collection of (mental) capabilities that enable an entity

1. to **solve** specific (difficult) **problems** (or to learn how to solve them),  
—**solve problems**—
2. to **master the unknown**: to act successfully in known, unknown and dynamic environments (perception, planning, agency, etc.)  
—**master the unknown**—

# AI – Own working definition

## Def.: Artificial Intelligence

A science of computational technologies that are developed to achieve and explain **intelligent** behavior in machines.

## Def.: Intelligence

Collection of (mental) capabilities that enable an entity

1. to **solve** specific (difficult) **problems** (or to learn how to solve them),  
—**solve problems**—
2. to **master the unknown**: to act successfully in known, unknown and dynamic environments (perception, planning, agency, etc.)  
—**master the unknown**—
3. to **explore abstract theories** and to **reason rationally**, avoiding contradictions,  
—**be rational & abstract**—

# AI – Own working definition

## Def.: Artificial Intelligence

A science of computational technologies that are developed to achieve and explain **intelligent** behavior in machines.

## Def.: Intelligence

Collection of (mental) capabilities that enable an entity

1. to **solve** specific (difficult) **problems** (or to learn how to solve them),  
—**solve problems**—
2. to **master the unknown**: to act successfully in known, unknown and dynamic environments (perception, planning, agency, etc.)  
—**master the unknown**—
3. to **explore abstract theories** and to **reason rationally**, avoiding contradictions,  
—**be rational & abstract**—
4. to **self-reflect** and to align one's own reasoning with overriding goals and standards, and  
—**self-reflect**—

# AI – Own working definition

## Def.: Artificial Intelligence

A science of computational technologies that are developed to achieve and explain **intelligent** behavior in machines.

## Def.: Intelligence

Collection of (mental) capabilities that enable an entity

1. to **solve** specific (difficult) **problems** (or to learn how to solve them),  
—**solve problems**—
2. to **master the unknown**: to act successfully in known, unknown and dynamic environments (perception, planning, agency, etc.)  
—**master the unknown**—
3. to **explore abstract theories** and to **reason rationally**, avoiding contradictions,  
—**be rational & abstract**—
4. to **self-reflect** and to align one's own reasoning with overriding goals and standards, and  
—**self-reflect**—
5. to **interact socially** with other entities and to adapt own goals and norms to those of a community (for a greater good).  
—**be social**—

# AI – Own working definition

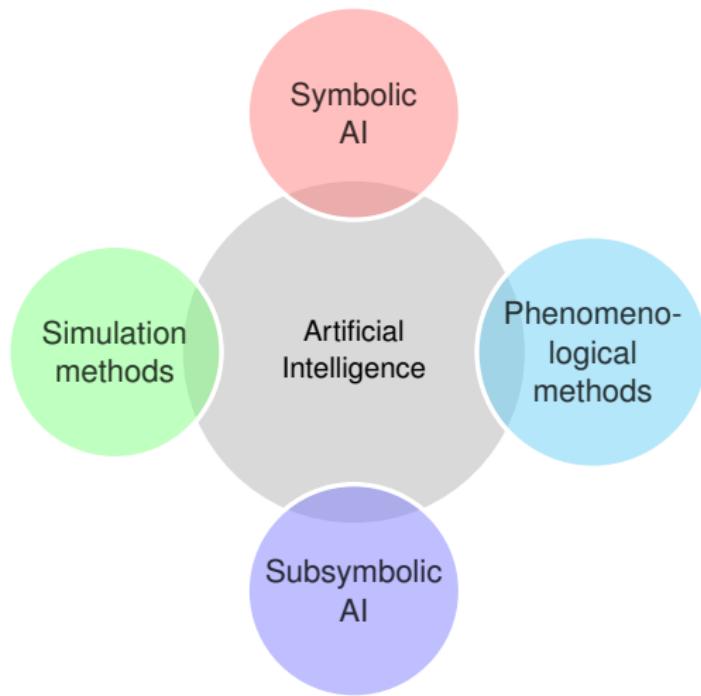
The problem (in my opinion) is not an emerging superintelligence!

The problem is:

- ▶ use of »severely limited« AI technology (lacking competencies 3-5)
  - ... within increasingly complex system environments
  - ... to serve most critical applications
  - ... in (semi-)autonomous mode

1. to **solve** specific (difficult) **problems** (or to learn how to solve them),  
—**solve problems**—
2. to **master the unknown**: to act successfully in known, unknown and dynamic environments (perception, planning, agency, etc.)  
—**master the unknown**—
3. to **explore abstract theories** and to **reason rationally**, avoiding contradictions,  
—**be rational & abstract**—
4. to **self-reflect** and to align one's own reasoning with overriding goals and standards, and  
—**self-reflect**—
5. to **interact socially** with other entities and to adapt own goals and norms to those of a community (for a greater good).  
—**be social**—

# AI – The notion (unfortunately) has changed



- ▶ Knowledge Representation
- ▶ Logical Reasoning
- ▶ Planning
- ▶ Multi-Agent Systems
- ▶ Search
- ▶ ...
- ▶ Robotics
- ▶ Pattern Recognition & Machine Learning
- ▶ Neural Networks

Current (very narrow/limited) notion of AI: **Neural Networks & ML**

# ML not tied to Neural Networks

LOGIC JOURNAL  
of the  
**IGPL**

Article Navigation

## Automatic Learning of Proof Methods in Proof Planning

Mateja Jamnik, Manfred Kerber, Martin Pollet, Christoph Benzmüller

*Logic Journal of the IGPL*, Volume 11, Issue 6, November 2003, Pages 647–673,

<https://doi.org/10.1093/jigpal/11.6.647>

**Published:** 01 November 2003

PDF

See e.g. also: Inductive Logic Programming

# Neural Networks & Machine Learning

Source: Fjodor van Veen, asimovinstitute.org, 2016

A mostly complete chart of

## Neural Networks

©2016 Fjodor van Veen - asimovinstitute.org

○ Backfed Input Cell

○ Input Cell

△ Noisy Input Cell

● Hidden Cell

○ Probabilistic Hidden Cell

△ Spiking Hidden Cell

○ Output Cell

○ Match Input Output Cell

● Recurrent Cell

○ Memory Cell

△ Different Memory Cell

● Kernel

○ Convolution or Pool

Perceptron (P)



Feed Forward (FF)



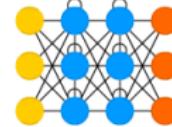
Radial Basis Network (RBF)



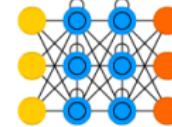
Deep Feed Forward (DFF)



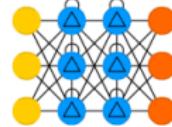
Recurrent Neural Network (RNN)



Long / Short Term Memory (LSTM)



Gated Recurrent Unit (GRU)



Auto Encoder (AE)



Variational AE (VAE)



Denoising AE (DAE)



Sparse AE (SAE)



# Neural Networks & Machine Learning

Source: Fjodor van Veen, asimovinstitute.org, 2016

Markov Chain (MC)



Hopfield Network (HN)



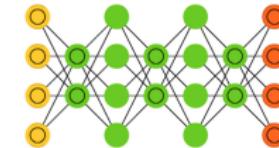
Boltzmann Machine (BM)



Restricted BM (RBM)



Deep Belief Network (DBN)



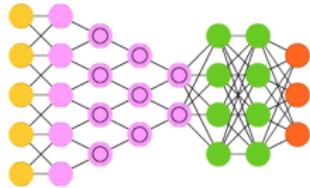
FF

GRU

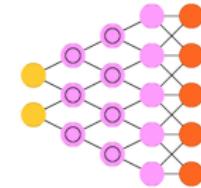
GRU

AE

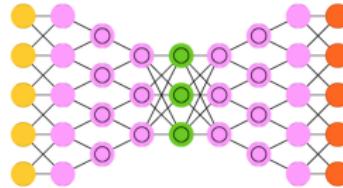
Deep Convolutional Network (DCN)



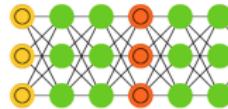
Deconvolutional Network (DN)



Deep Convolutional Inverse Graphics Network (DCIGN)



Generative Adversarial Network (GAN)



Liquid State Machine (LSM)



Extreme Learning Machine (ELM)



Echo State Network (ESN)



Deep Residual Network (DRN)



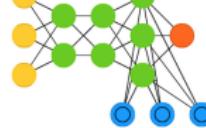
Kohonen Network (KN)



Support Vector Machine (SVM)



Neural Turing Machine (NTM)



# Neural Networks & Machine Learning

Source: Fjodor van Veen, asimovinstitute.org, 2016

Markov Chain (MC)



Hopfield Network (HN)



Boltzmann Machine (BM)



Restricted BM (RBM)



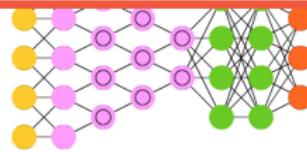
Deep Belief Network (DBN)



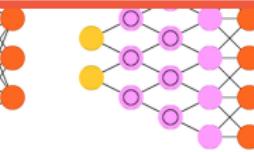
FF

## Sub-/Non-Symbolic AI, since

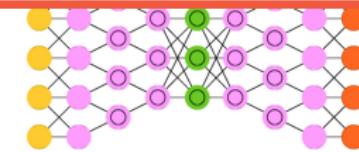
- ▶ nodes do not carry any semantic meaning



Generative Adversarial Network (GAN)

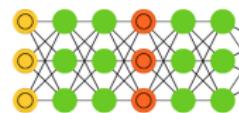


Liquid State Machine (LSM)

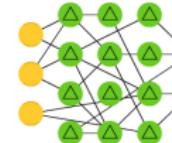


Extreme Learning Machine (ELM)

GRU



Deep Residual Network (DRN)



Kohonen Network (KN)



Support Vector Machine (SVM)



Neural Turing Machine (NTM)

AE

GRU

# Neural Networks & Machine Learning

Source: Fjodor van Veen, asimovinstitute.org, 2016

Markov Chain (MC)



Hopfield Network (HN)



Boltzmann Machine (BM)



Restricted BM (RBM)



Deep Belief Network (DBN)



FF

## Sub-/Non-Symbolic AI, since

- ▶ nodes do not carry any semantic meaning

GRU

### Pros:

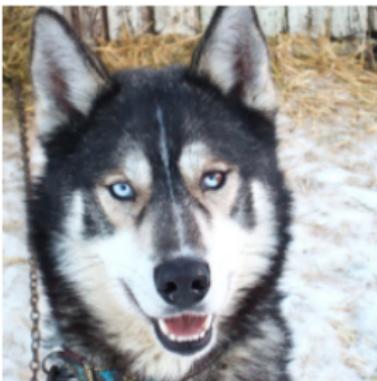
- ▶ robust & very strong in specific domains

### Cons:

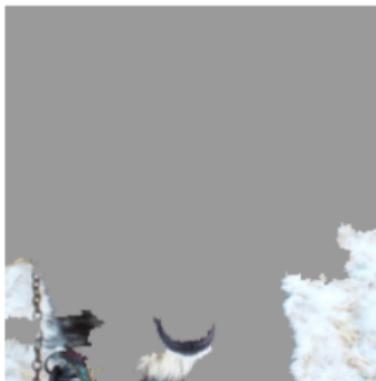
- ▶ data-intensive training, requires substantial human expertise, intransparent, bias, adversarial attacks



# Neural Networks & Machine Learning



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Source: Ribeiro et al., arXiv:1602.04938

**Explanation of black box AI systems: often without sense!!  
How could »opening this black box« possibly create trust?**



# Symbolic AI – Example: Logical Reasoning

$A \cup B ::= \dots$

$A \cap B ::= \dots$

$A \subseteq B ::= \dots$

$A = B ::= \dots$

$\dots$

$\dots$

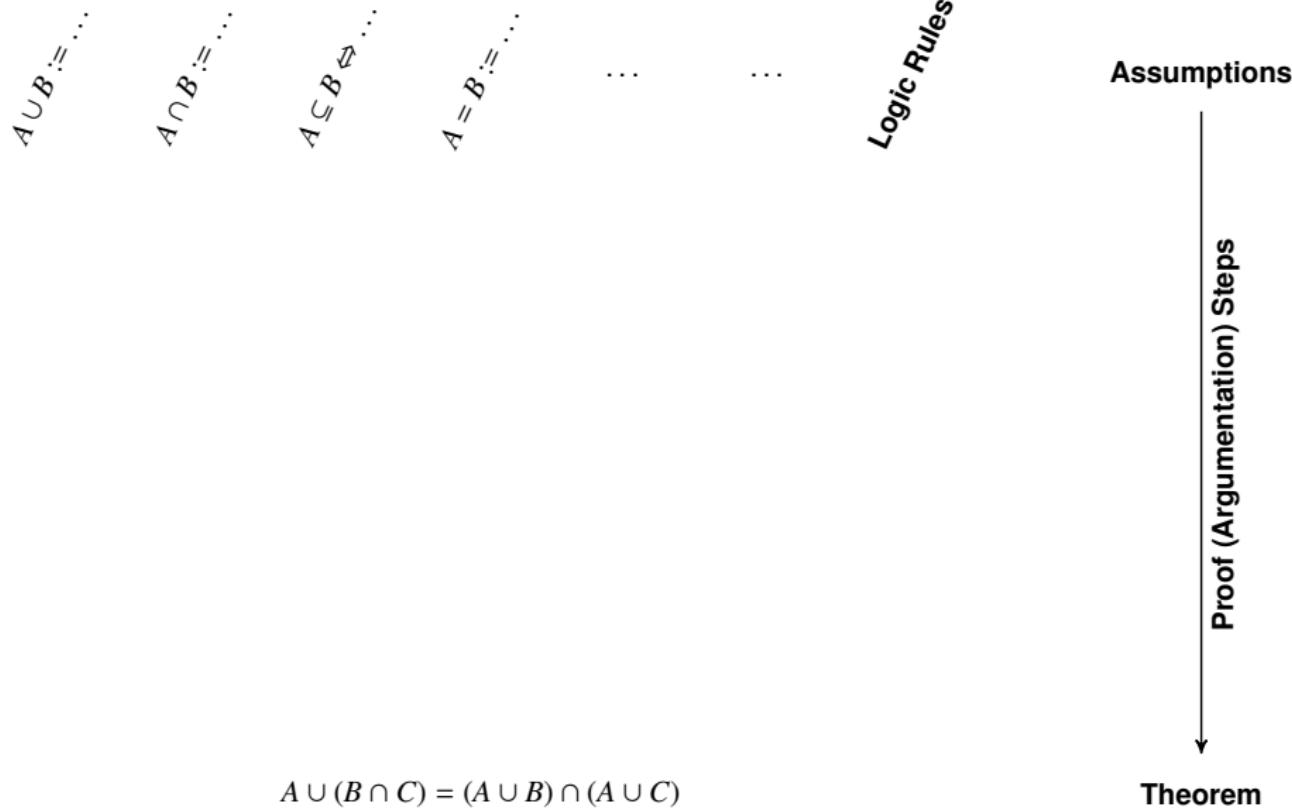
*Logic Rules*

**Assumptions**

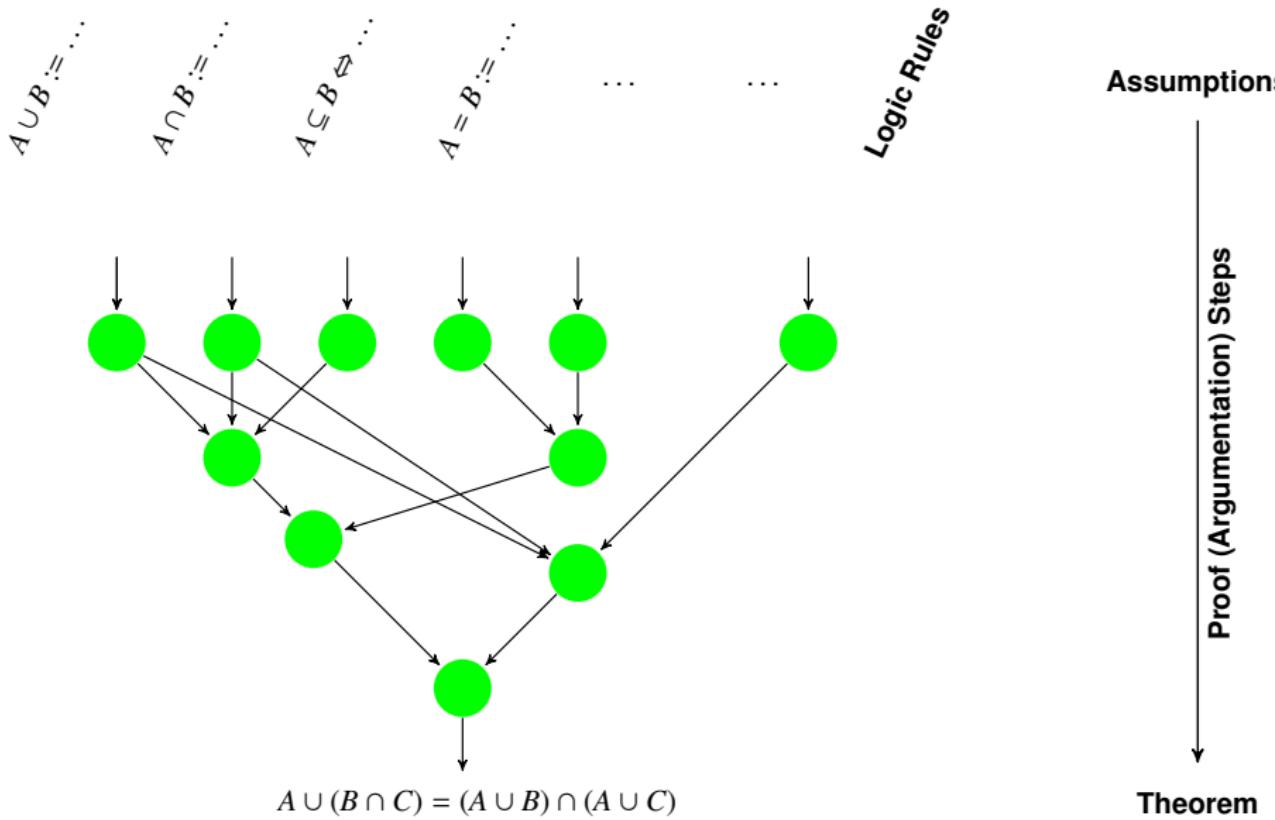
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

**Theorem**

# Symbolic AI – Example: Logical Reasoning



# Symbolic AI – Example: Logical Reasoning



# Symbolic AI – Example: Logical Reasoning



## Automated (Symbolic) Reasoners:

AI systems...

... which automatically search for such proof arguments

Own tools: Leo-prover family (Leo-I/II/III)

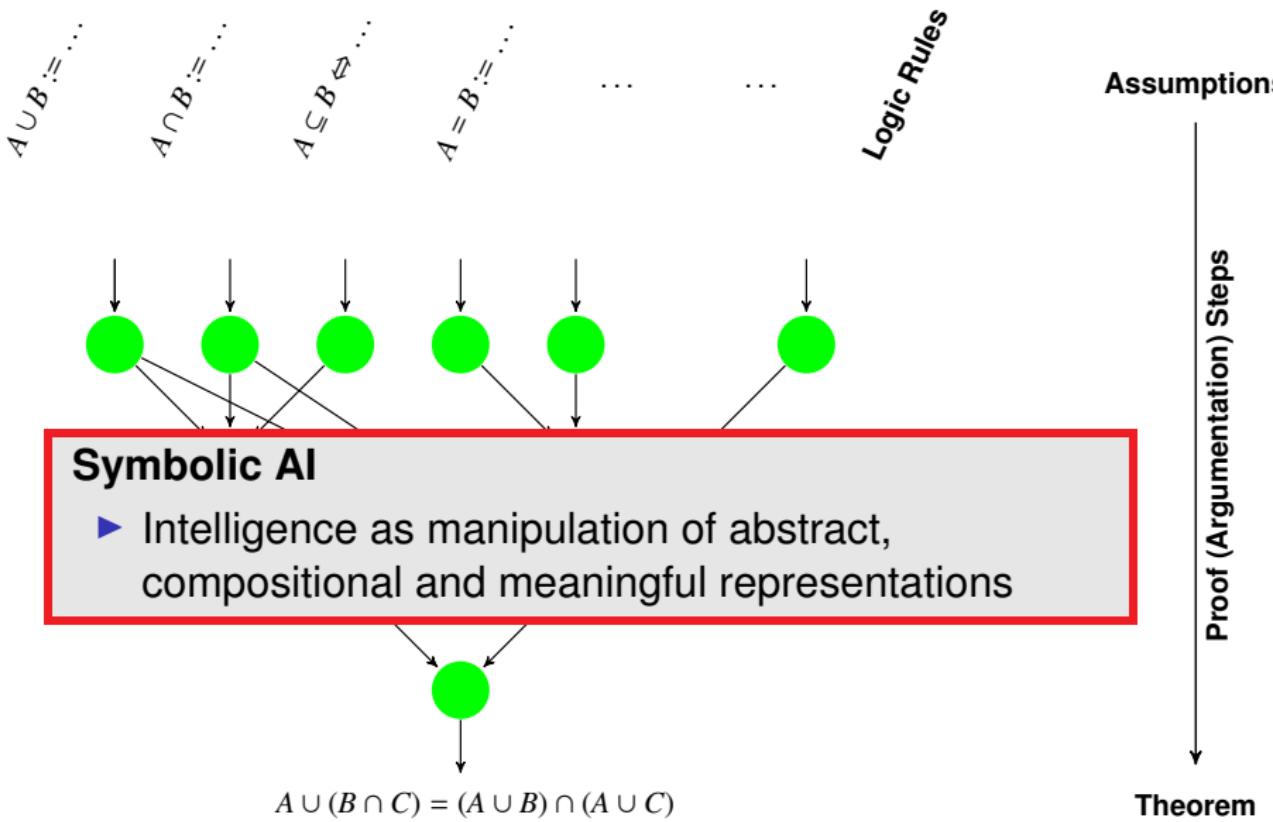
Many applications, including

**Rational Argumentation (Philosophy, Ethics&Law)**

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

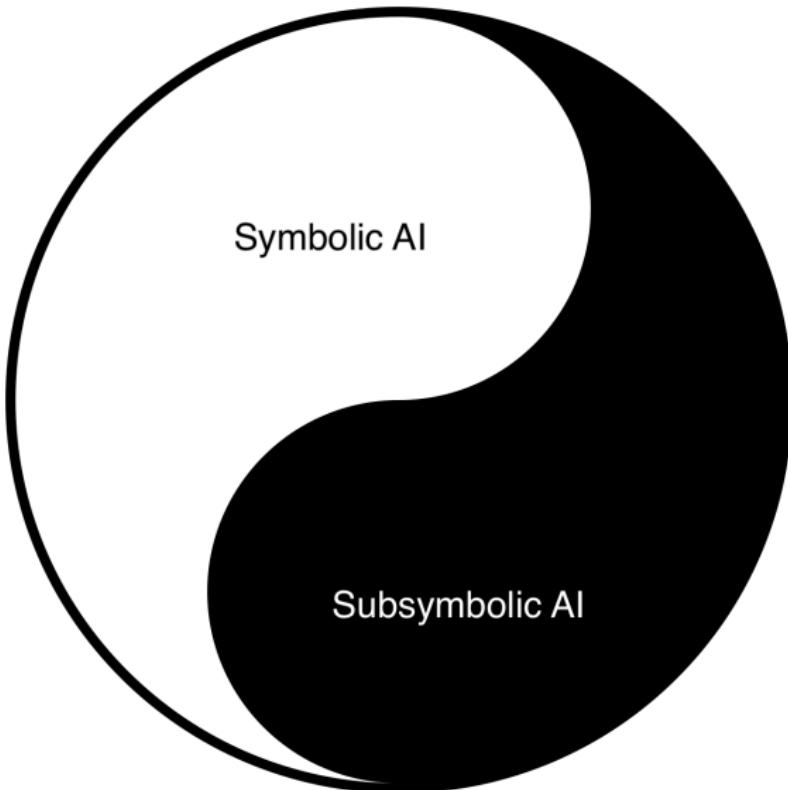
Theorem

# Symbolic AI – Example: Logical Reasoning



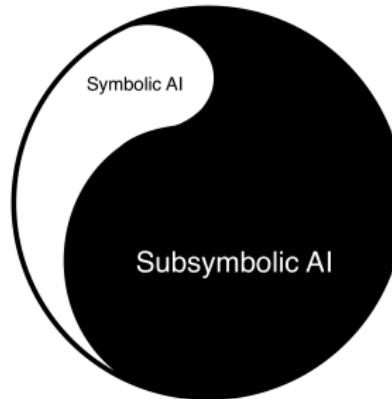
# **Own Position**

# Yin and Yang of AI — Own Position



There are many recent calls for a “**Neuro-Symbolic AI**”  
(but actually this is an “old hat”, isn’t it?)

# Yin and Yang of AI — Unhealthy Hype!

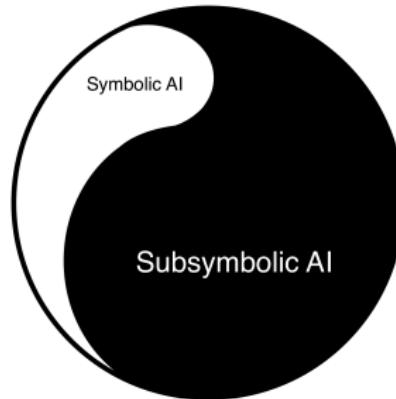


## Examples

- ▶ ...
- ▶ AlphaGo und AlphaZero: world champion in Chess and Go

(**subsymbolic KI**)

# Yin and Yang of AI — Unhealthy Hype!



## Examples

- ▶ ...
- ▶ AlphaGo und AlphaZero: world champion in Chess and Go

(**subsymbolic KI**)

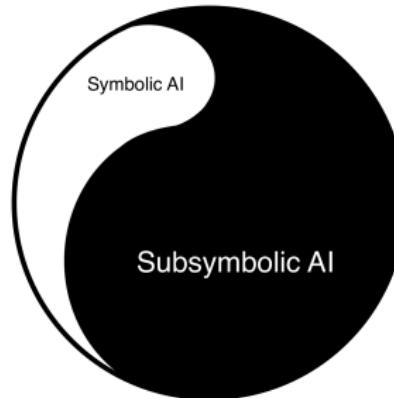


# Yin and Yang of AI — Unhealthy Hype!

## Examples

- ▶ ...
- ▶ SAT-Solver:  
solution of open  
maths problems

(symbolic AI)



## Examples

- ▶ ...
- ▶ AlphaGo und  
AlphaZero: world  
champion in Chess  
and Go

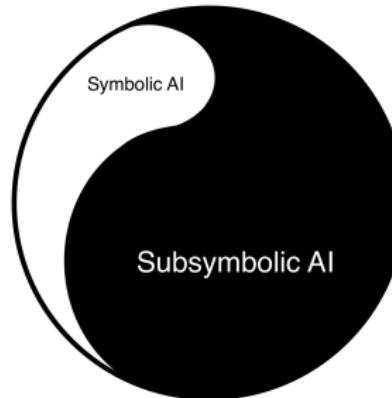
(subsymbolic KI)



# Yin and Yang of AI — Unhealthy Hype!

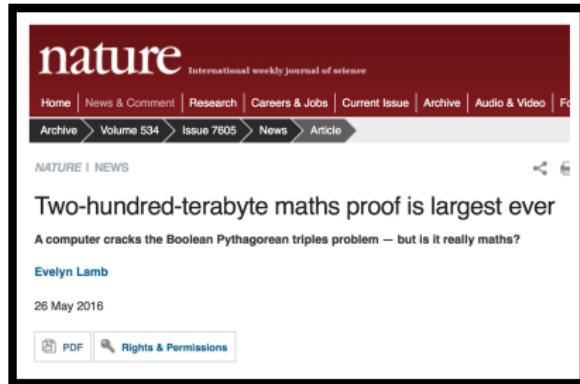
## Examples

- ▶ ...
- ▶ SAT-Solver:  
solution of open  
maths problems  
  
**(symbolic AI)**



## Examples

- ▶ ...
- ▶ AlphaGo und  
AlphaZero: world  
champion in Chess  
and Go  
  
**(subsymbolic KI)**



**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | Fo

Archive > Volume 534 > Issue 7605 > News > Article

NATURE | NEWS

Two-hundred-terabyte maths proof is largest ever

A computer cracks the Boolean Pythagorean triples problem — but is it really maths?

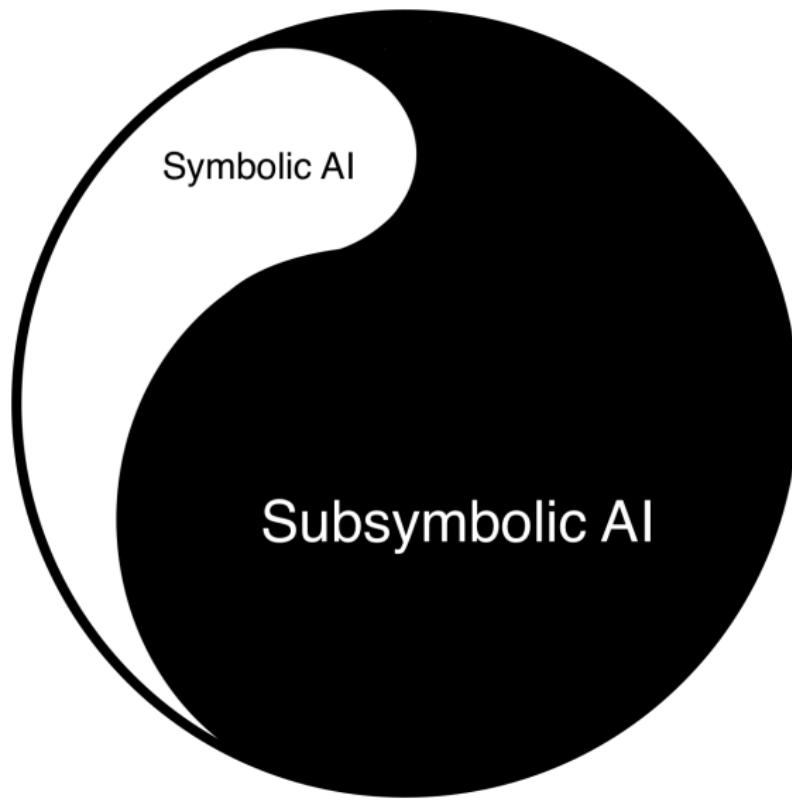
Evelyn Lamb

26 May 2016

PDF Rights & Permissions



# **Yin and Yang of AI — Unhealthy Hype!**



**While many sharp minds are already rethinking!**

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

Causalities

Abstraction

Precise Reasoning

Explain-/Verifiability

...

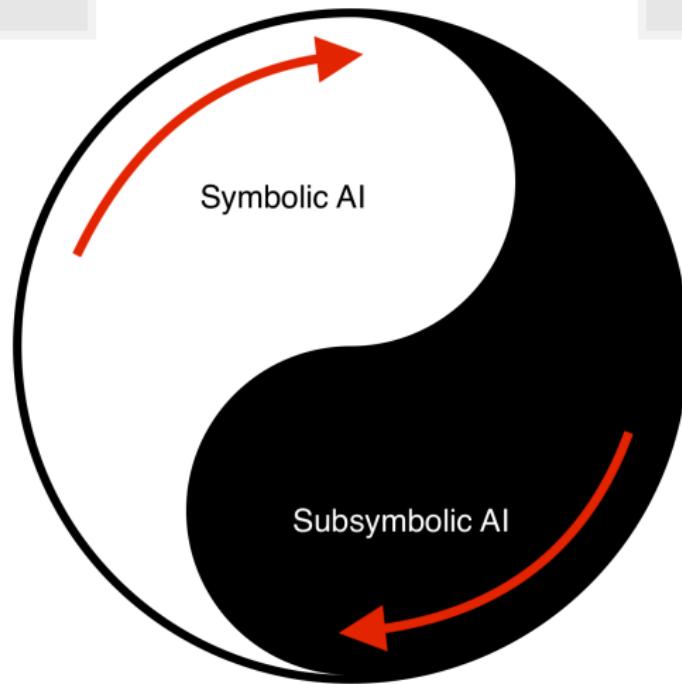
Correlations

Patterns

Robustness

Learning

...



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

Causalities

Abstraction

Precise Reasoning

Explain-/Verifiability

...

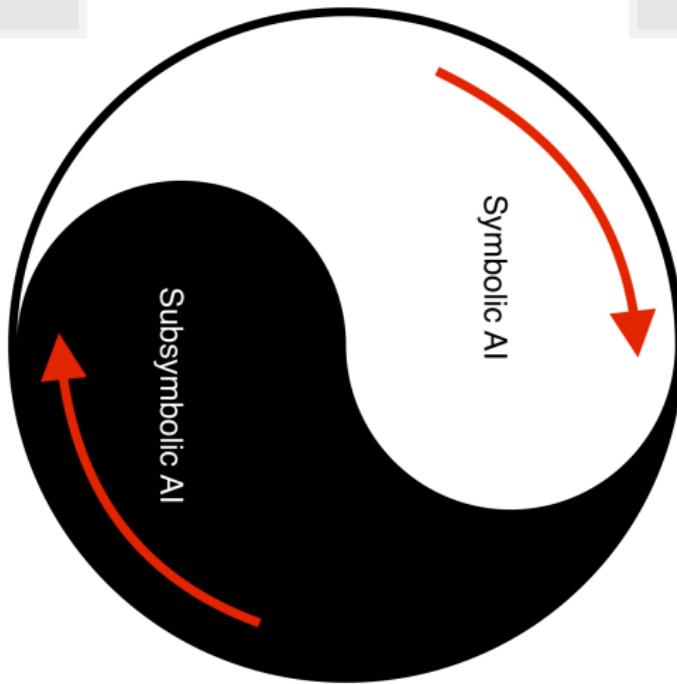
Correlations

Patterns

Robustness

Learning

...



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

Causalities

Abstraction

Precise Reasoning

Explain-/Verifiability

...

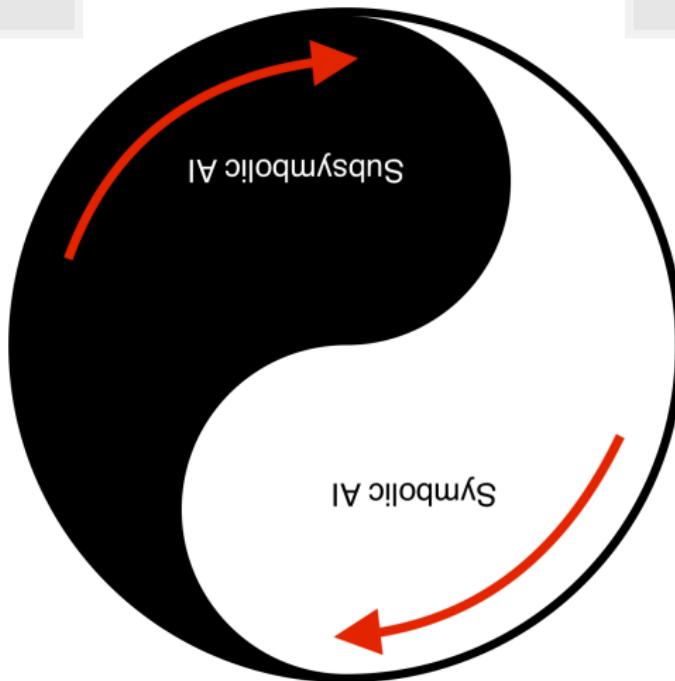
Correlations

Patterns

Robustness

Learning

...



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

Causalities

Abstraction

Precise Reasoning

Explain-/Verifiability

...

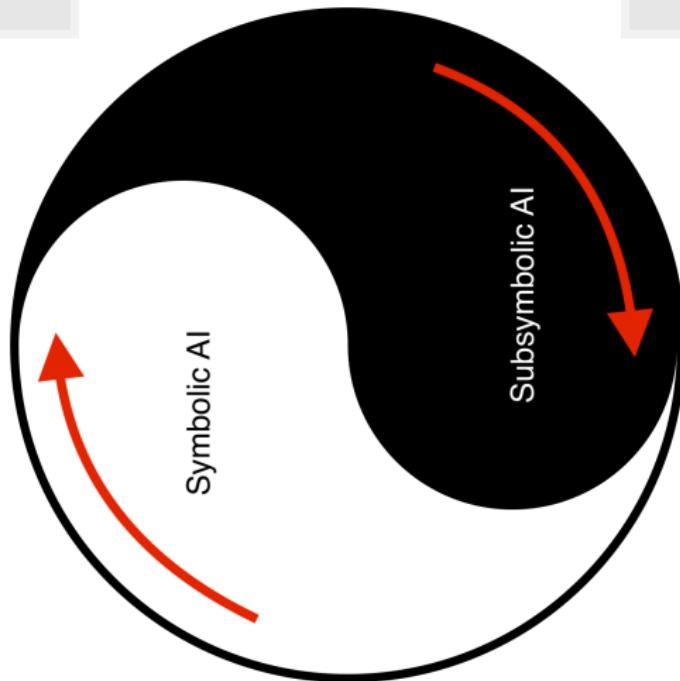
Correlations

Patterns

Robustness

Learning

...



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

Causalities

Abstraction

Precise Reasoning

Explain-/Verifiability

...

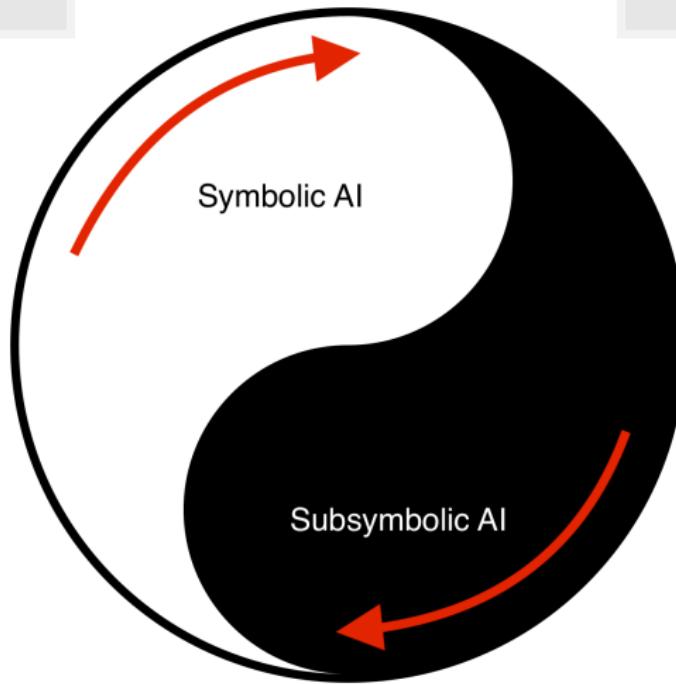
Correlations

Patterns

Robustness

Learning

...



# **How to create trust & how to control?**

# How to create trust?



# How to create trust?



English title: "Rebel Without a Cause"

- ▶ Do our current AI systems know what they are doing?
- ▶ Do we know what we are doing when we entrust such AI systems with increasingly critical decisions?
- ▶ Is normative directionlessness and unpredictability the core character of future AI systems?
- ▶ Why should we trust such systems?

# How to create trust?



English title: "Rebel Without a Cause"

- ▶ Do our current AI systems know what they are doing?
- ▶ Do we know what we are doing when we entrust such AI systems with increasingly critical decisions?
- ▶ Is normative directionlessness and unpredictability the core character of future AI systems?
- ▶ Why should we trust such systems?

# How to create trust?



English title: "Rebel Without a Cause"

- ▶ Do our current AI systems know what they are doing?
- ▶ Do we know what we are doing when we entrust such AI systems with increasingly critical decisions?
- ▶ Is normative directionlessness and unpredictability the core character of future AI systems?
- ▶ Why should we trust such systems?

# How to create trust?



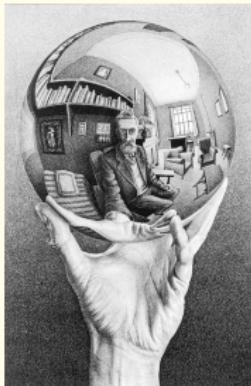
English title: "Rebel Without a Cause"

- ▶ Do our current AI systems know what they are doing?
- ▶ Do we know what we are doing when we entrust such AI systems with increasingly critical decisions?
- ▶ Is normative directionlessness and unpredictability the core character of future AI systems?
- ▶ Why should we trust such systems?

# How to create trust?



- ▶ Do our current AI systems know what they are doing?
- ▶ Do we know what we are doing



"Hand with Reflecting Sphere (1935)" by M.C. ESCHER.

- ▶ We need mature/responsible citizens capable of
- ▶ ... introspection & self-reflection
- ▶ ... engaging in **rational dialogues**
- ▶ Do erratic AI systems fit in this picture?
- ▶ Are they ready for societal integration?
- ▶ First, we need to create trust!



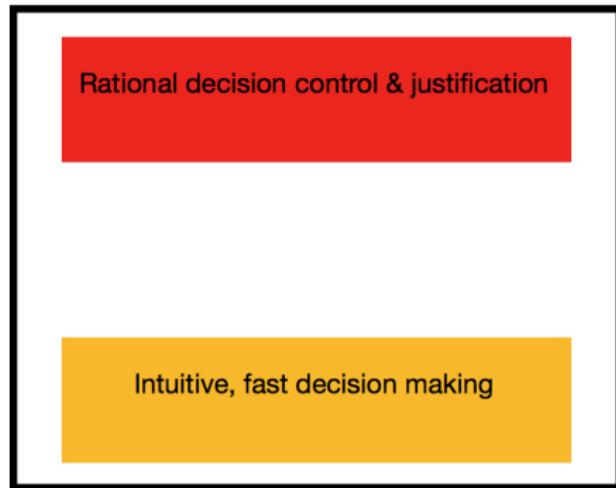
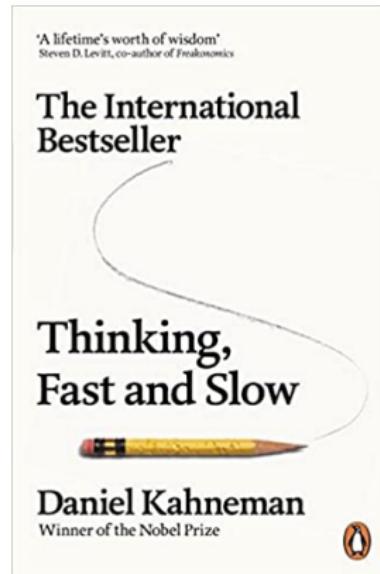
Humanoid "Sophia"



English title: "Rebel Without a Cause"

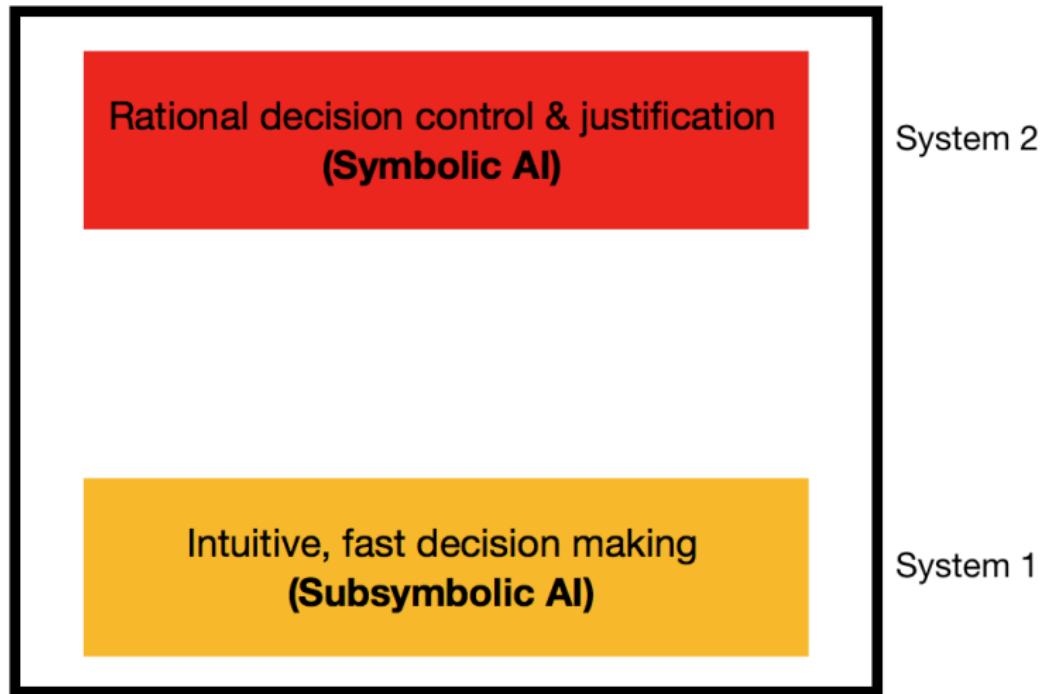
systems?

# How to create trust? How to control?

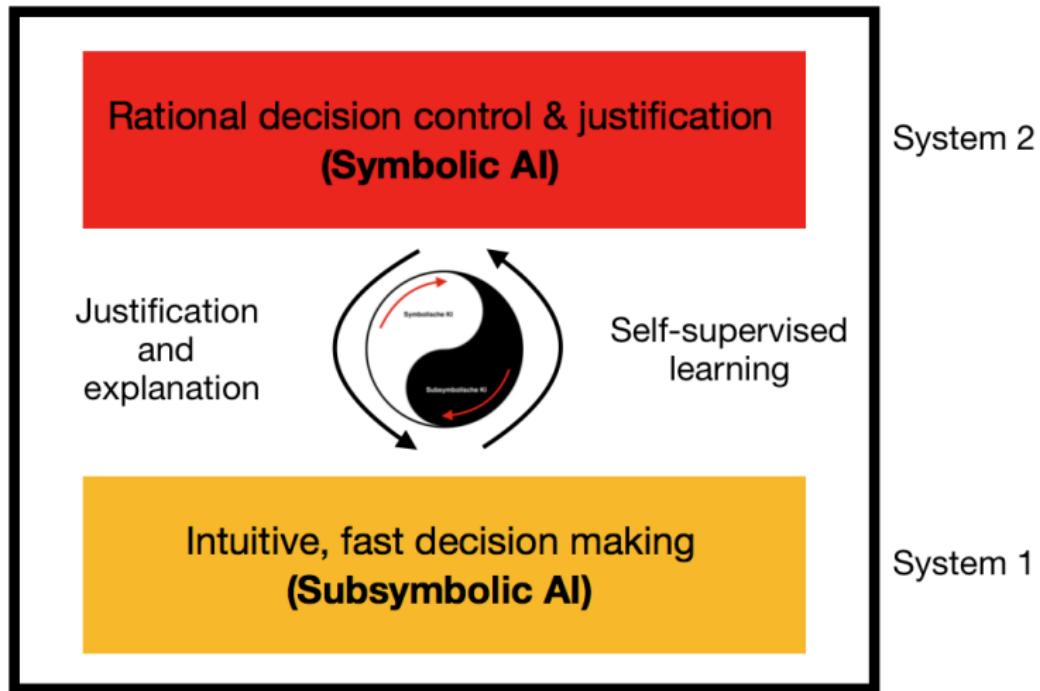


See e.g. also: **J. Haidt, The Emotional Dog and its Rational Tail, 2001**

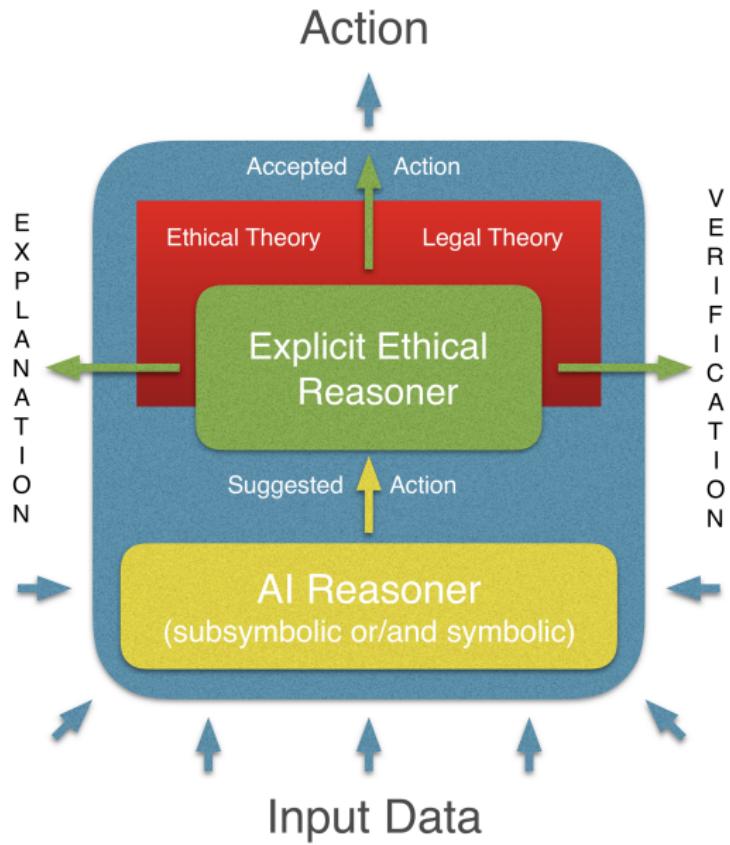
# How to create trust? How to control?



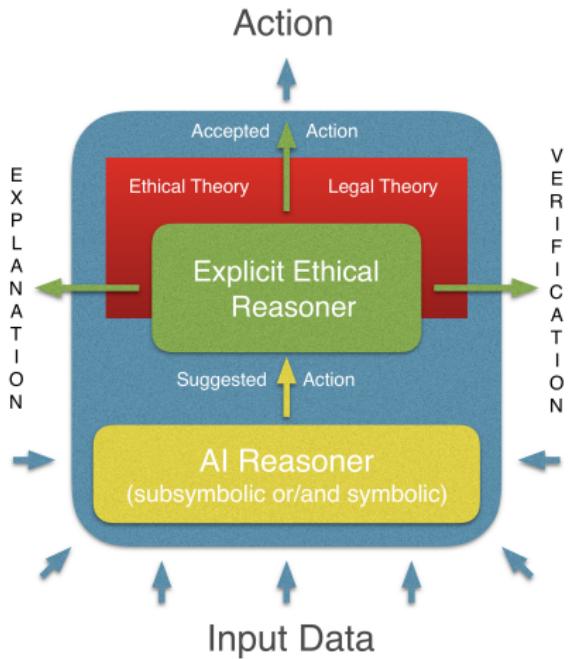
# How to create trust? How to control?



# How to control? Pseudo-Ethical AI Agent



# How to control? Pseudo-Ethical AI Agent



## Related Works

- ▶ 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]

Addresses demands for explainability and verifiability:  
⇒ but not at the level of AI black box systems!

[German Conference on Artificial Intelligence \(Künstliche Intelligenz\)](#)

... KI 2020: [KI 2020: Advances in Artificial Intelligence pp 251-258](#) | [Cite as](#)

# Reasonable Machines: A Research Manifesto

Authors

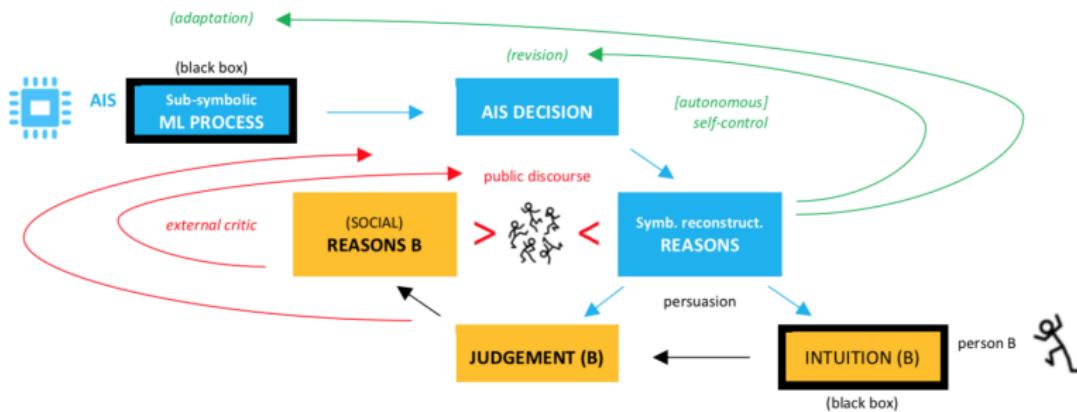
[Authors and affiliations](#)

---

Christoph Benzmüller  , Bertram Lomfeld



## Artificial “Social Reasoning Model”



**Building trust into AI systems through rational communication of »reasons«**

Important:

- ▶ these reasons may actually be independent of the original motivational impulse to act
- ▶ should be possible to avoid opening the black-box, since this can increase vulnerability

# Recent Research Focus: Experiments in expressive, symbolic KR&R in the fields of ethics and law



Designing normative theories for ethical and legal reasoning: LogiKEY framework, methodology, and tool support \*

Christoph Benzmueller<sup>1, 4, 5, 6</sup>, Xavier Parent<sup>2, 3</sup>, Leendert van der Torre<sup>4, 5, 6</sup>

Show more ▾

<https://doi.org/10.1016/j.artint.2020.103348>

Get rights and content



Encoding Legal Balancing: Automating an Abstract Ethico-Legal Value Ontology in Preference Logic  
Christoph Benzmueller, David Fuenmayer, Bertram Lomfeld

ECAL 2020  
G.D. Giacomo et al. (Eds.)  
© 2020. The authors and IOS Press.  
This is an open access publication with Open Access by IOS Press and distributed under the terms  
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).  
doi:10.3233/PALI200045

## Normative Reasoning with Expressive Logic Combinations

David Fuenmayer<sup>1</sup> and Christoph Benzmueller<sup>2</sup>



Search

Download book

Pacific Rim International Conference on Artificial Intelligence

CLAR 2020: Logic and Argumentation pp 104-115 | Cite as

Computer-Supported Analysis of Arguments in Climate Engineering  
Harnessing Higher-Order (Meta-)Logic to Represent and Reason with Complex Ethical Theories

Authors Authors and affiliations

David Fuenmayer , Christoph Benzmueller

Authors Authors and affiliations

David Fuenmayer , Christoph Benzmueller

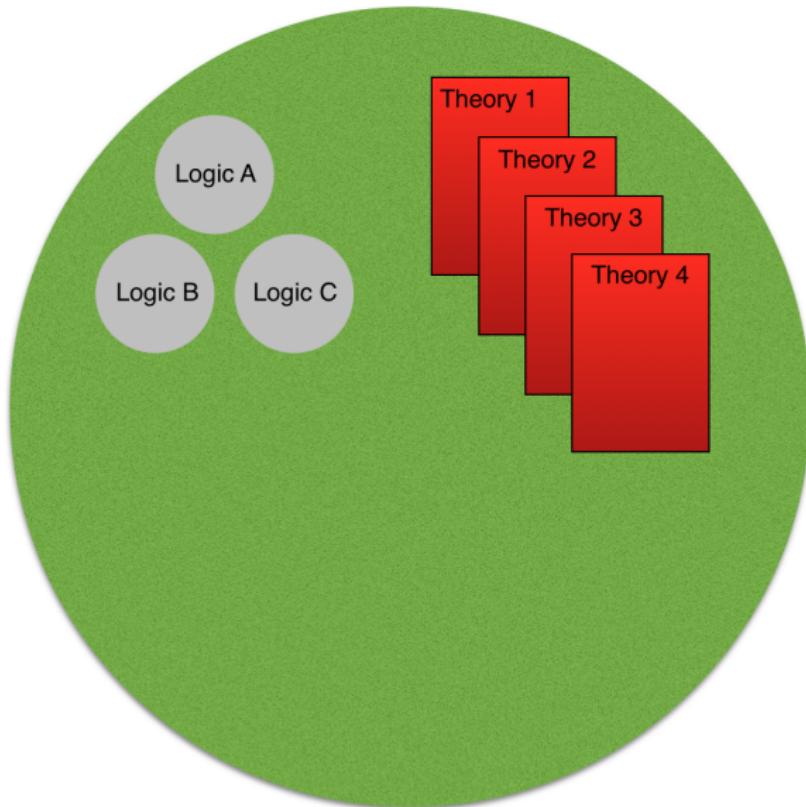
(some recent papers)



# Normative Reasoning Experimentation Framework

(proposed in 2018 DEON keynote and much further developed since then)

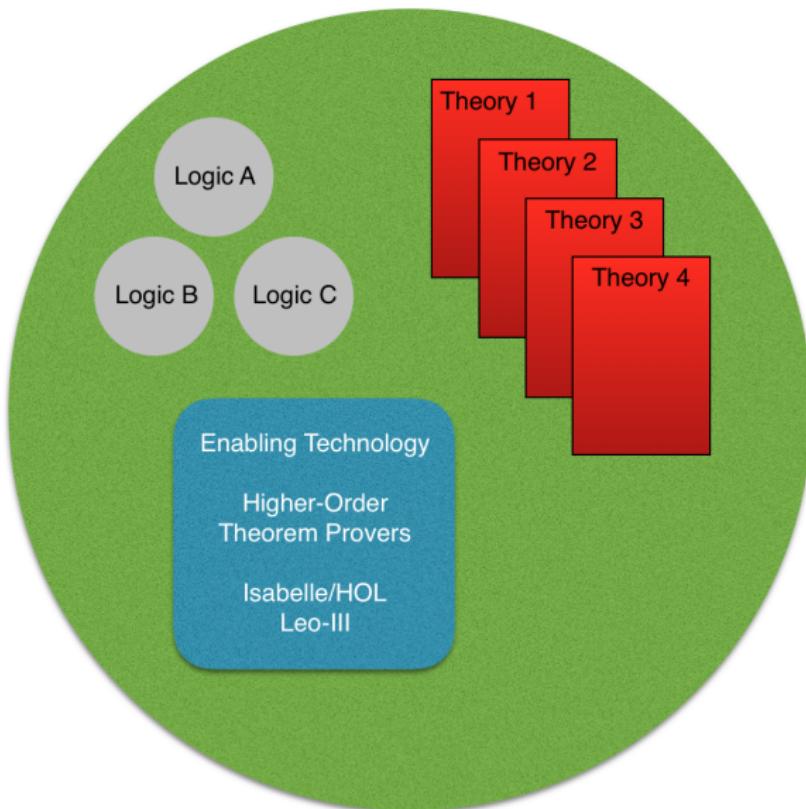
[AIJ (2020) vol. 287], [Data in Brief (2020) vol. 33], [www.logikey.org](http://www.logikey.org)



# Normative Reasoning Experimentation Framework

(proposed in 2018 DEON keynote and much further developed since then)

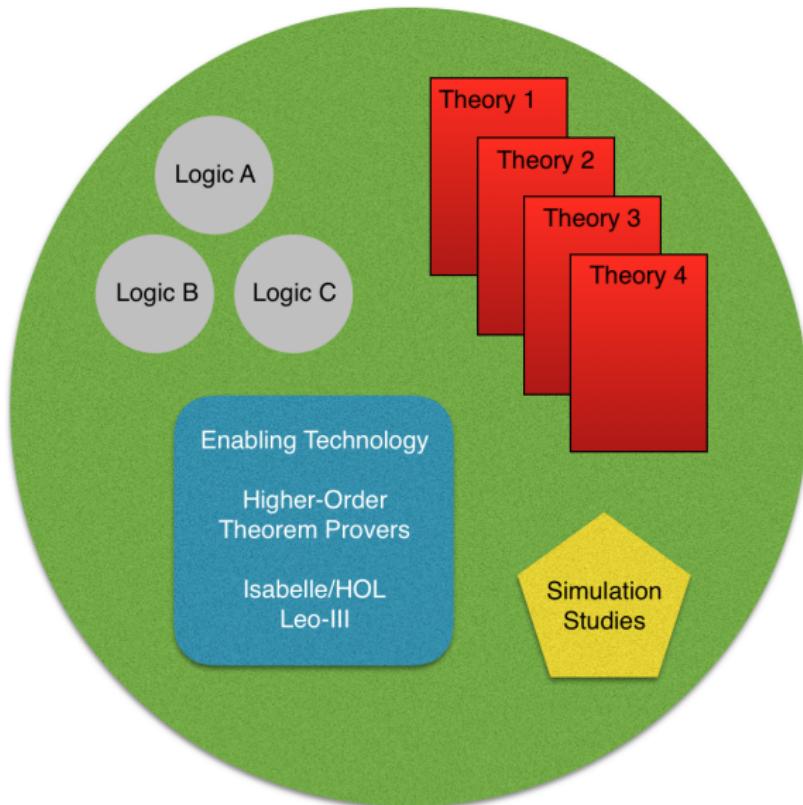
[AIJ (2020) vol. 287], [Data in Brief (2020) vol. 33], [www.logikey.org](http://www.logikey.org)



# Normative Reasoning Experimentation Framework

(proposed in 2018 DEON keynote and much further developed since then)

[AIJ (2020) vol. 287], [Data in Brief (2020) vol. 33], [www.logikey.org](http://www.logikey.org)



# Universal (Meta-)Logical Reasoning in HOL

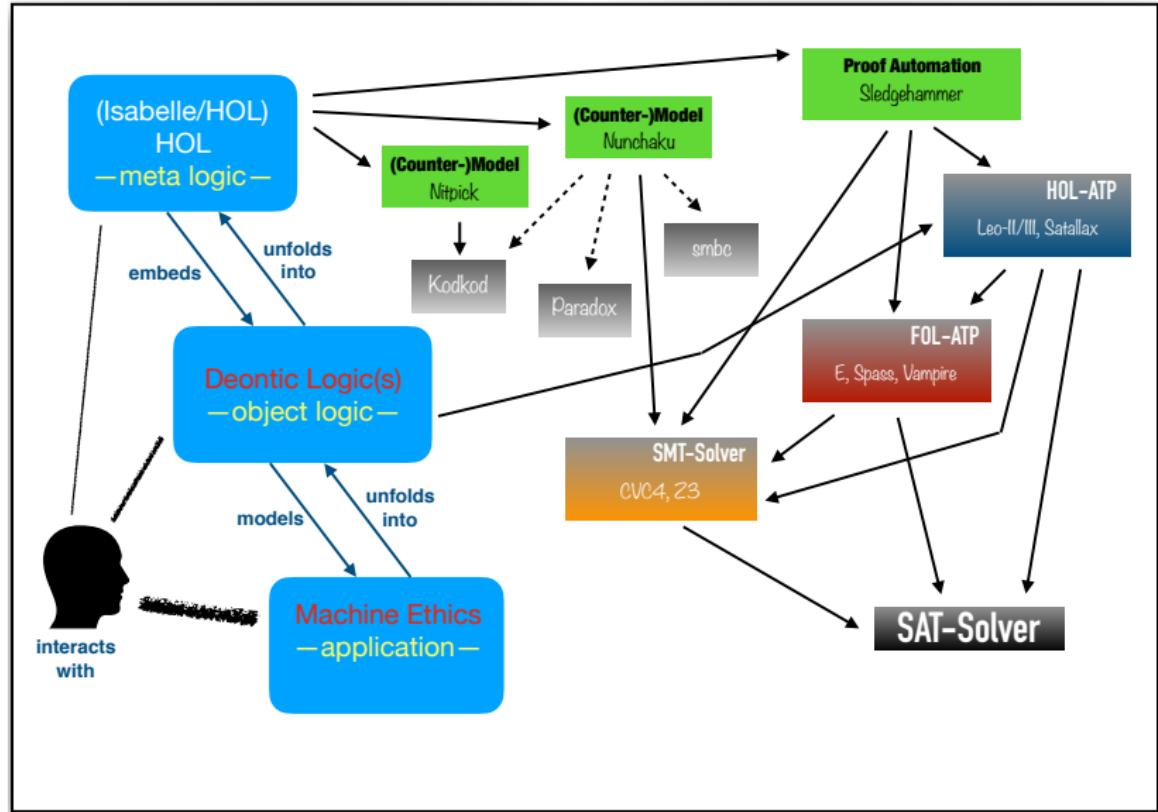
[SciCompProgr (2019) vol. 172]



How to Tame the Logic Zoo?

# Universal (Meta-)Logical Reasoning in HOL

[SciCompProgr (2019) vol. 172]





Artificial Intelligence

Volume 287, October 2020, 103348



# Designing normative theories for ethical and legal reasoning: LogIKEY framework, methodology, and tool support ☆

Christoph Benzmüller<sup>b, a</sup>  , Xavier Parent<sup>a</sup> , Leendert van der Torre<sup>a, c</sup> 

Show more ▾

<https://doi.org/10.1016/j.artint.2020.103348>

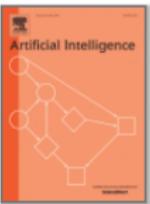
[Get rights and content](#)



ELSEVIER

# Artificial Intelligence

Volume 287, October 2020, 103348



Use Cases

Domain Knowledge

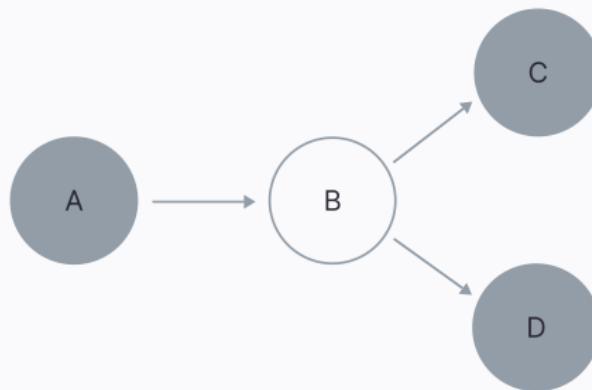
(Combinations of) Object Logic(s)

Meta-Logic HOL

LogKEy Methodology

# LogKEy and Abstract Argumentation?

[CLAR (2020)], [CLAR (2018)], [MBR (2018)]

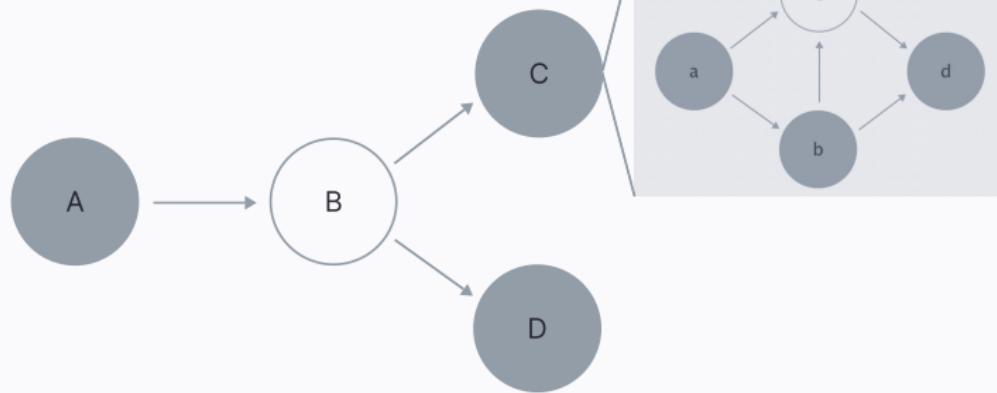


Pitch

**Research goal:** Show that the LogKEy framework supports  
pluralistic semantics

# LogKEy and Abstract Argumentation?

[CLAR (2020)], [CLAR (2018)], [MBR (2018)]

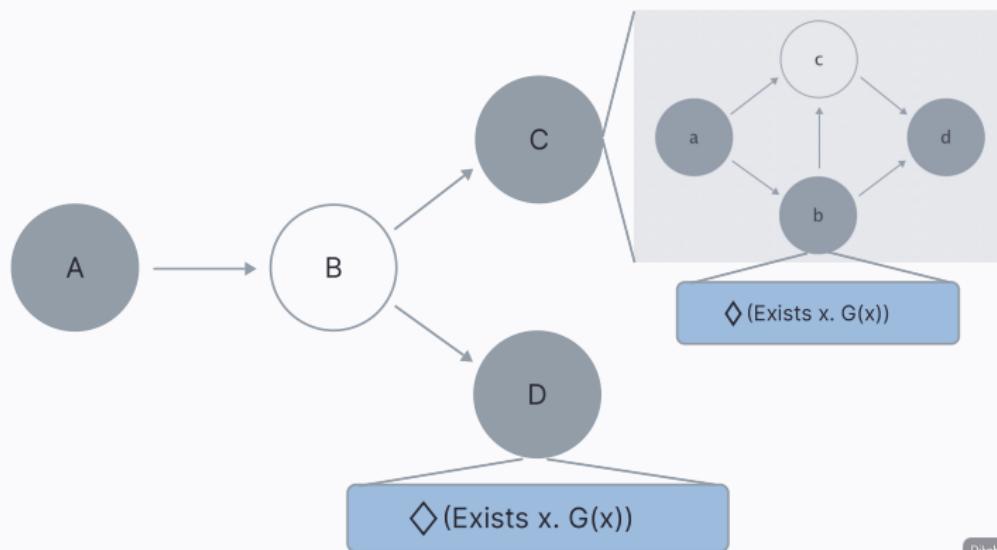


Pitch

**Research goal:** Show that the LogKEy framework supports  
pluralistic semantics | nested arguments

# LogKEy and Abstract Argumentation?

[CLAR (2020)], [CLAR (2018)], [MBR (2018)]



**Research goal:** Show that the LogKEy framework supports  
pluralistic semantics | nested arguments | nodes with complex formulas

# LogiKEY and Legal Balancing?

For example, Wild Animal Court Cases — [MLR@KR (2020)]



*Post, a fox hunter, was chasing a fox through public land when Pierson came across the fox and, knowing it was being chased, killed the fox and took it away. Post sued Pierson for damages against his possession of the fox. Post argued that giving chase to the fox was sufficient to establish possession.*

- ▶ A local court first ruled in favour of Post
- ▶ Pierson appealed, decision was changed

# LogiKEY and Legal Balancing?

For example, Wild Animal Court Cases — [MLR@KR (2020)]



*Post, a fox hunter, was chasing a fox through public land when Pierson came across the fox and, knowing it was being chased, killed the fox and took it away. Post sued Pierson for damages against his possession of the fox. Post argued that giving chase to the fox was sufficient to establish possession.*



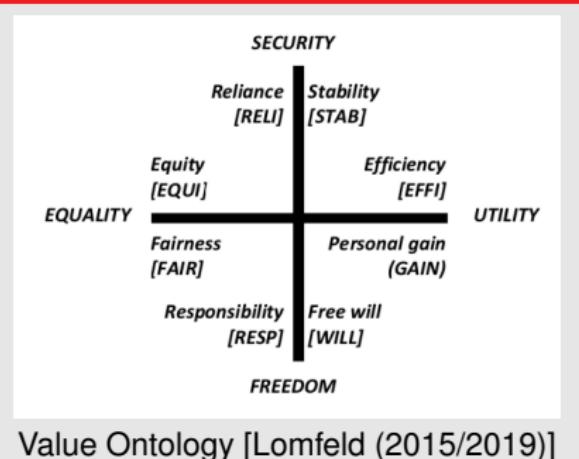
*Chester, a parrot owned by the ASPCA (animal shelter), escaped and was recaptured by Conti. The ASPCA found this out and reclaimed Chester from Conti.*

- ▶ Here the court ruled in favour of the ASPCA

- ▶ A local court first ruled in favour of Post
- ▶ Pierson appealed, decision was changed

# LogIKEy and Legal Balancing?

For example, Wild Animal Court Cases — [MLR@KR (2020)]



Post, a fox hunter, was chasing a fox on public land when Pierson saw him. Post had been hunting the fox and, knowing it was being chased, killed the fox and took it away. Post sued Pierson for damages against his possession of the fox. Post argued that giving chase to the fox was sufficient to establish possession.

- ▶ A local court first ruled in favour of Post
- ▶ Pierson appealed, decision was changed

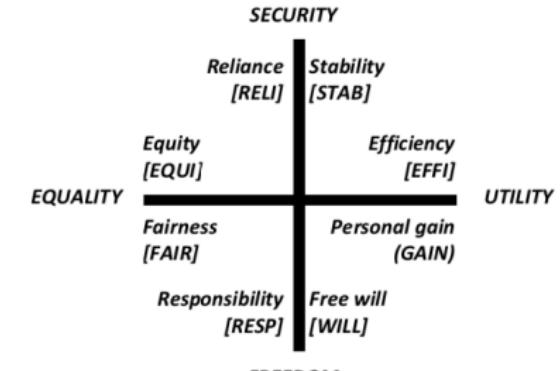


led by the ASPCA (animal welfare organization) was recaptured by Conti. The court and reclaimed Chester

- ▶ Here the court ruled in favour of the ASPCA

# LogIKey and Legal Balancing?

For example, Wild Animal Court Cases — [MLR@KR (2020)]



Value Ontology [Lomfeld (2015/2019)]

*Post, a fox hunter, was chasing a fox on public land when Pierson saw it. Post had given chase to the fox and, knowing it was being chased, killed the fox and took it away. Post sued Pierson for damages against his possession of the fox. Post argued that giving chase to the fox was sufficient to establish possession.*

- ▶ A local court first ruled in favour of Post
- ▶ Pierson appealed, decision was changed
- ▶ The decision in favour of Pierson implies: legal **STABILITY** > **WILL** to possess

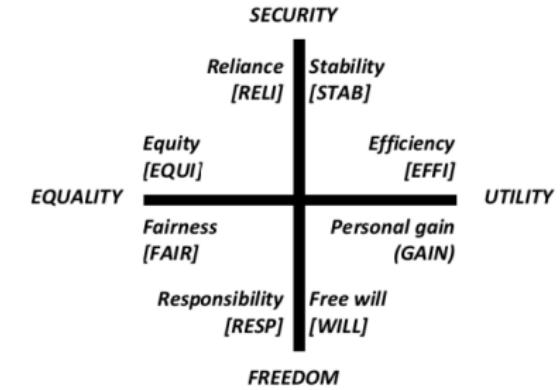


*led by the ASPCA (animal welfare organization) was recaptured by Conti. Conti had been out and reclaimed Chester*

- ▶ Here the court ruled in favour of the ASPCA
- ▶ For **domestic animals** value preference **STABILITY** > **WILL** does not apply
- ▶ For a **domestic animal** it is sufficient that owner did not give up the **RESPonsibility** for its maintenance
- ▶ **RESP** together with ASPCA's **RELIance** in the parrot's property > Conti's corporal possession (**STAB**) of the animal

# LogIKEy and Legal Balancing?

For example, Wild Animal Court Cases — [MLR@KR (2020)]



Post, a fox hunter, was claiming public land when Pierson and, knowing it was his, went against his possessory rights that giving chase established possession

Value Ontology [Lomfeld (2015/2019)]

R1: "[appAnimal → (STAB<sup>p</sup> ↼v STAB<sup>d</sup>)]" and  
R2: "[appWildAnimal → (WILL<sup>x-1</sup> ↼v STAB<sup>x</sup>)]" and  
R3: "[appDomAnimal → (STAB<sup>x-1</sup> ↼v RELI<sup>x</sup>⊕RESP<sup>x</sup>)]"

**abbreviation** "Pierson\_facts ≡ [Fox α ∧ (FreeRoaming α) ∧ (¬Pet α) ∧ Pursue p α ∧ (¬Pursue d α) ∧ Capture d α]"

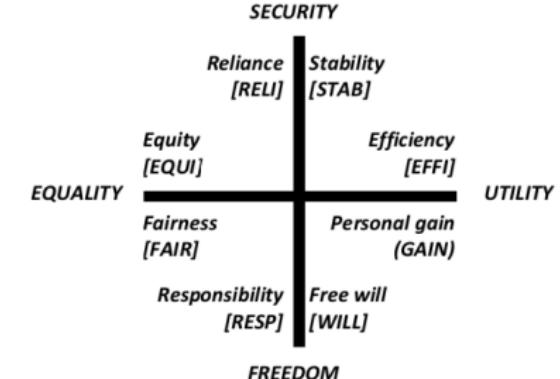
**theorem assumes** Pierson\_facts **shows** "[For p ↼ For d]"  
by (metis assms CW1 CW2 W6 W8 ForAx R2 F1 other.simps(2) rBR)

- ▶ A local court ruled in favour of Pierson
- ▶ Pierson appealed to the court of appeals
- ▶ The decision in favour of Pierson implies: legal **STABILITY** > **WILL** to possess

- ▶ RESP together with ASPCA's **RELIANCE** in the parrot's property > Conti's corporal possession (**STAB**) of the animal

# LogIKEy and Legal Balancing?

For example, Wild Animal Court Cases — [MLR@KR (2020)]



Post, a fox hunter, was cited for hunting on public land when Pierson, a local landowner, saw him and, knowing it was illegal, stopped him and took it away. Pierson sued Post for trespass against his possession of the land, which was established by giving chase and capturing the animal.

- ▶ A local court ruled in favor of Pierson.
- ▶ Pierson appealed to the state supreme court.
- ▶ The decision upheld the legal **STABILITY**.

R1: " $\text{appAnimal} \rightarrow (\text{STAB}^d \prec_v \text{STAB}^d)$ " and  
R2: " $\text{appWildAnimal} \rightarrow (\text{WILL}^{x-1} \prec_v \text{STAB}^x)$ " and  
R3: " $\text{appDomAnimal} \rightarrow (\text{STAB}^{x-1} \prec_v \text{RELI}^x \oplus \text{RESP}^x)$ "

**abbreviation** "Pierson\_facts"  $\equiv$  [Fox  $\alpha \wedge$  (FreeRoaming  $\alpha \wedge$  ( $\neg$ Pet  $\alpha \wedge$  Pursue p  $\alpha \wedge$  ( $\neg$ Pursue d  $\alpha \wedge$  Capture d  $\alpha)$ ))]"

**theorem assumes** Pierson\_facts **shows** "[For p  $\prec$  For d]"  
by (metis assms CW1 CW2 W6 W8 ForAx R2 F1 other.simps(2) rBR)

**abbreviation** "ASPCA\_facts"  $\equiv$  [Parrot  $\alpha \wedge$  Pet  $\alpha \wedge$  Care p  $\alpha \wedge$  Prop p  $\alpha \wedge$  ( $\neg$ Prop d  $\alpha \wedge$  Capture d  $\alpha)$ ]"

**lemma aux:** **assumes** ASPCA\_facts **shows** "[ $(\text{STAB}^d \prec_v \text{RELI}^p \oplus \text{RESP}^p)$ ]"  
using CW1 CW2 W7 assms R3 by fastforce

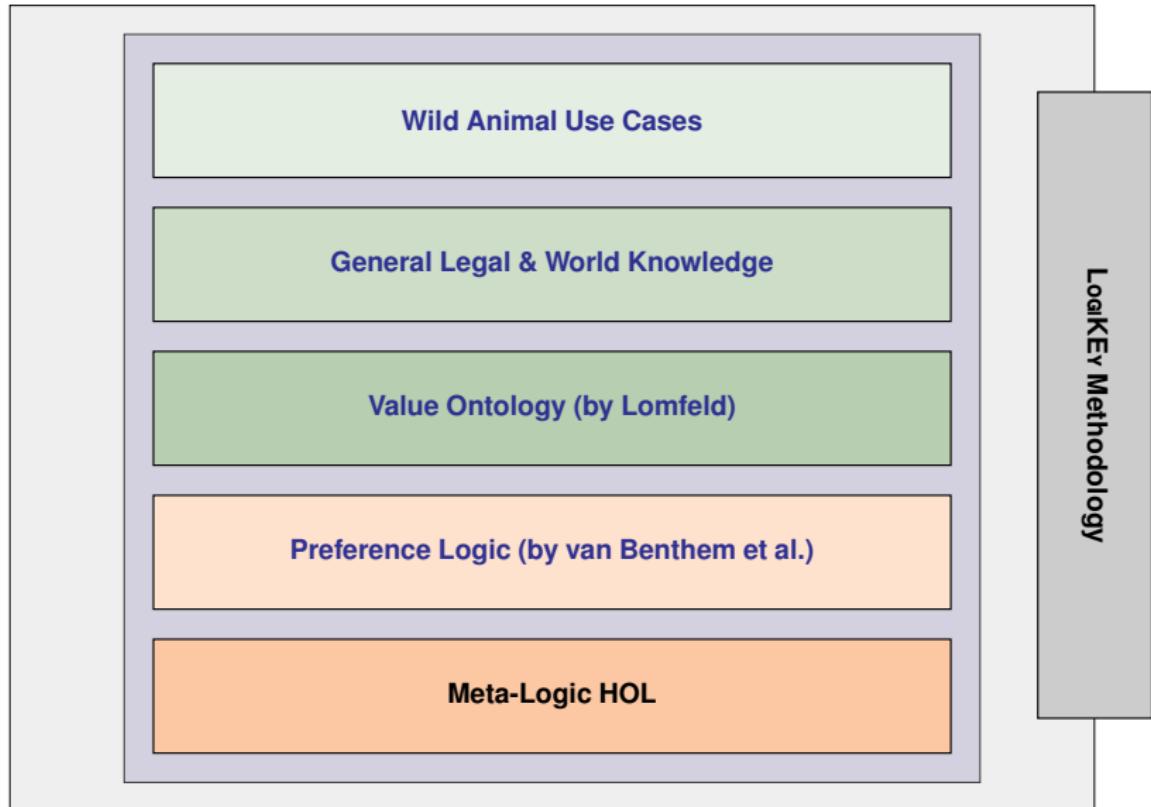
**theorem assumes** ASPCA\_facts **shows** "[For d  $\prec$  For p]"  
using assms aux CW5 ForAx F3 other.simps(1) rBR by metis

jurisdiction of the ASPCA  
and the preference  
of the law apply  
and sufficient that  
ESPonsability for

's RELIance in  
is corporal  
animal

# LogIKEy and Legal Balancing

[MLR@KR (2020)]



# Conclusion – How to create trust?

Can post-hoc normative reasoning competencies prevent AI systems from going rogue?

- ▶ Opening AI black box systems – Really a good Idea?
  - ▶ transparency (in case of imperfect systems) will have limited impact on trust
  - ▶ could even be seen as an invitation for adversarial attacks
- ▶ I instead argue for argumentation based harnesses
  - ▶ ... to control and justify decisions
  - ▶ they may be independent from encapsulated AI black boxes
  - ▶ they may benefit from opening the black box (inside the harness)
  - ▶ they may hide sensible information to the outside

