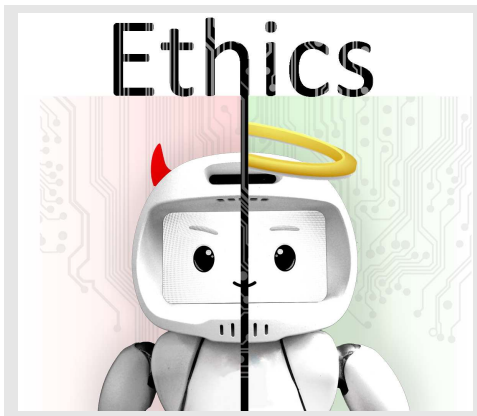


Ethico-legal Governance of AI Systems

Christoph Benz Müller

AI Systems Engineering (AISE), Universität Bamberg

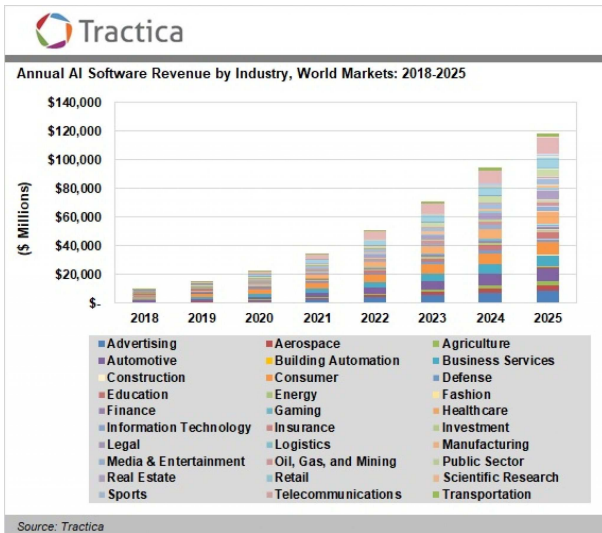


AI-Meetup, Dependable Intelligent Software Lab, 1. April 2022

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

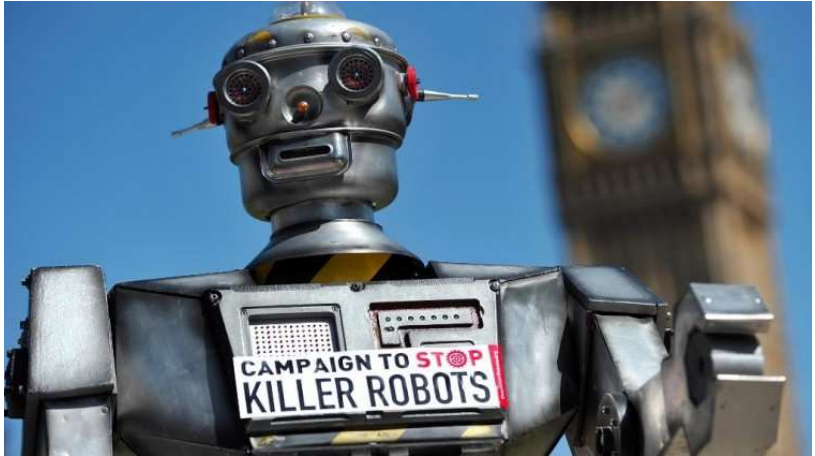
Bitkom: Digitalization to wipe out millions of jobs in Germany

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?

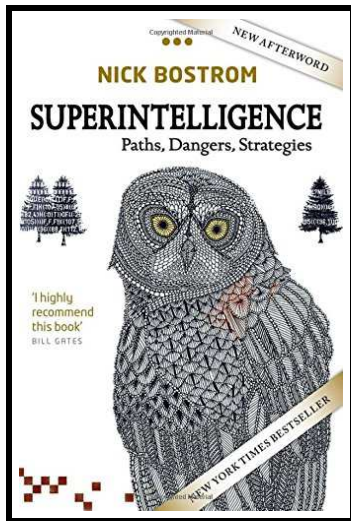
What is AI?



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

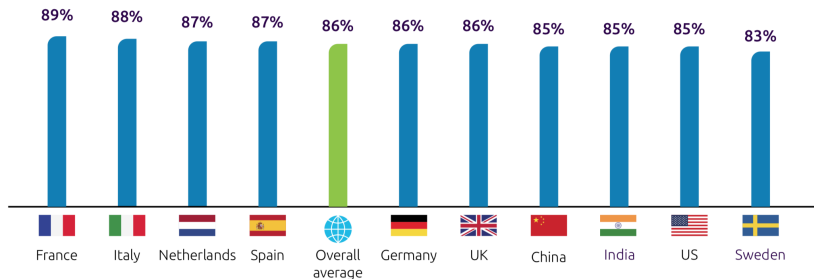
Source: MIT Technology Review, 2019

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)

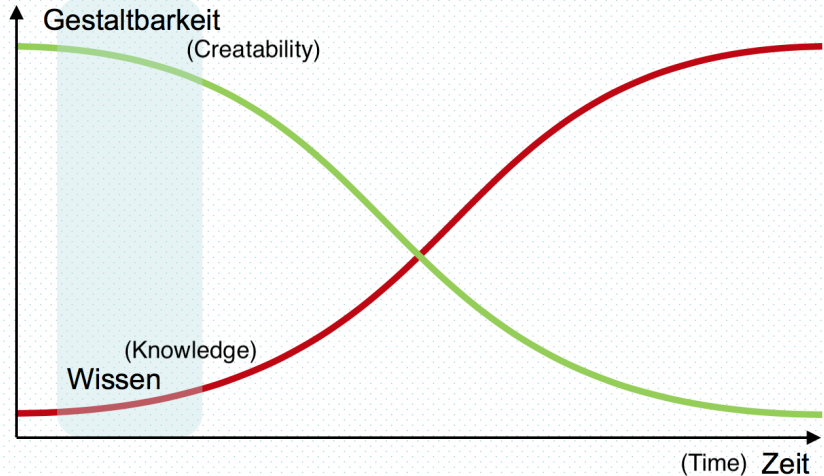


We presented over 40 cases where ethical issues could arise from the use of AI, to executives across sectors. We asked them whether they encountered these issues in the last 2-3 years.

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, strong in single domains

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, strong in single domains

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/Artificial

Limited 'smart' systems
single domain

- ▶ *AI is the science of making computers do things that would require human intelligence if they were done by a human. It is a biologically inspired approach to the study of how the mind works better. (Rich and Korth)*
- ▶ ...
- ▶ *Intelligence is a variety of abilities, including the ability to reason, to learn quickly and to solve problems, to acquire an academic skill, or to have the capability for common sense, for doing things, or 'figuring out' things.*

Strong AI

with (at least) general
human intelligence

... especially intelligent
computers to understand
methods that are

... moment, people are

... er things, involves the
... hend complex ideas,
... earning, a narrow
... er and deeper
... 'making sense' of



What is AI? – Many definitions, no consensus

Weak/Artificial

Limited 'smart' systems
single domain

Strong AI

Human (at least) general
intelligence

- ▶ AI is the science of making computers do things that would require human intelligence if they were done by a human. (Rich and Korth)
- ▶ The study of how to make computers do things better. (Rich and Korth)
- ▶ ...
- ▶ Intelligence is a capability to reason, to learn quickly and to acquire an academic skill, or a capability for common sense, or 'figuring out' things, or 'making sense' of things, or 'making sense' of things, or 'making sense' of things.

BRIEFING

EU Legislation in Progress

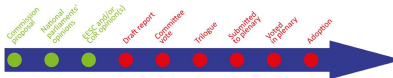


Artificial intelligence act

...

Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts

...



...

However, the Commission found that the **notion of an AI system** should be more clearly defined to ensure legal certainty, given that the determination of what an 'AI system' constitutes is crucial for the allocation of legal responsibilities under the new AI framework. The Commission therefore proposes to establish a legal definition of 'AI system' in EU law, which is largely based on a definition already used by the OECD.¹⁷ Article 3(1) of the draft act states that '**artificial intelligence system**' means:

...software that is developed with [specific] techniques and approaches [listed in Annex 1] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.¹⁸

Annex 1 of the proposal lays out a **list of techniques and approaches** that are used today to develop AI. Accordingly, the notion of 'AI system' would refer to a range of software-based technologies that encompasses '**machine learning**', '**logic and knowledge-based**' systems, and '**statistical**' approaches.¹⁹

...s, especially intelligent
computers to understand
methods that are

moment, people are

er things, involves the
hend complex ideas,
earning, a narrow
r and deeper
'making sense' of

What is AI? – Many definitions, no consensus

Weak/Approach
Limited 'smart' systems
single domain

Strong AI
General (at least) general
intelligence

BRIEFING

EU Legislation in Progress



Artificial intelligence act

. . .

Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts

► AI is the science
computer program
human intelligence

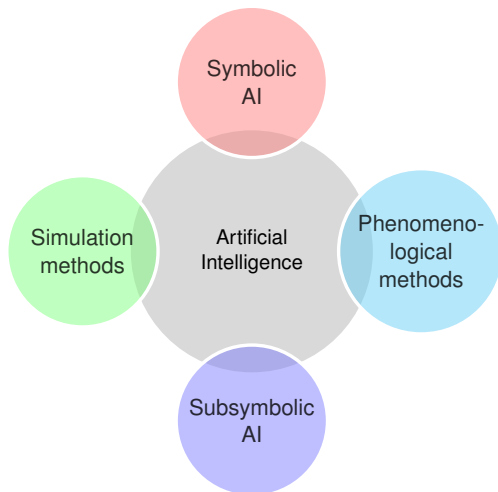
...s, especially intelligent
computers to understand
methods that are

However, the Commission found that the **notion of an AI system** should be more clearly defined to ensure legal certainty, given that the determination of what an 'AI system' constitutes is crucial for the allocation of legal responsibilities under the new AI framework. The Commission therefore proposes to establish a legal definition of 'AI system' in EU law, which is largely based on a definition already used by the OECD.¹⁷ Article 3(1) of the draft act states that '**artificial intelligence system**' means:

...software that is developed with [specific] techniques and approaches [listed in Annex 1] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.¹⁸

[Annex 1](#) of the proposal lays out a **list of techniques and approaches** that are used today to develop AI. Accordingly, the notion of 'AI system' would refer to a range of software-based technologies that encompasses '**machine learning**', '**logic and knowledge-based**' systems, and '**statistical**' approaches.¹⁹

AI – The currently hyped notion of AI ...



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

... strongly focuses on **Subsymbolic AI**

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!

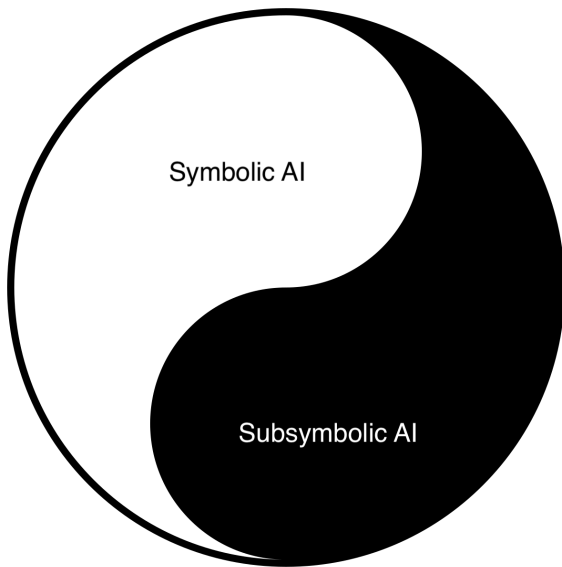
The problem is:

- ▶ 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**—

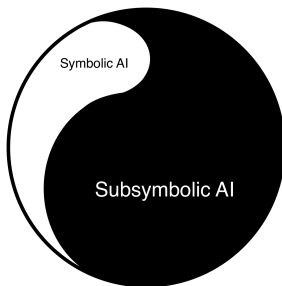
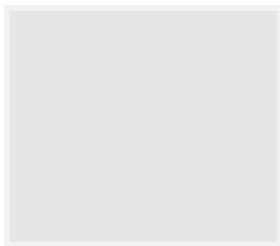
Own Position

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!

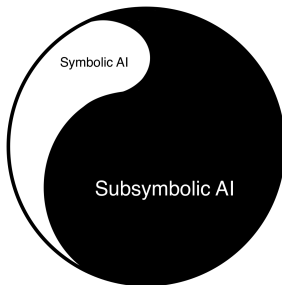
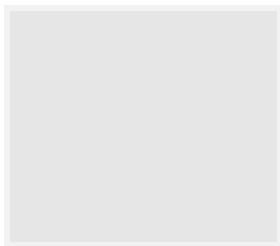


Examples

- ▶ ...
- ▶ 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)

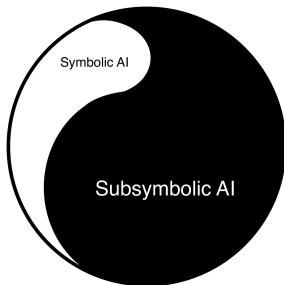


Yin and Yang of AI — Unhealthy Hype!

Examples

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

(symbolic AI)



Examples

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

(subsymbolic KI)

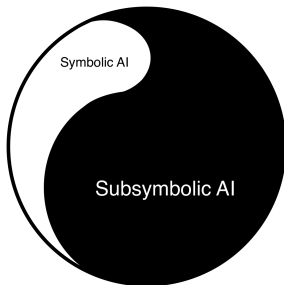


Yin and Yang of AI — Unhealthy Hype!

Examples

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

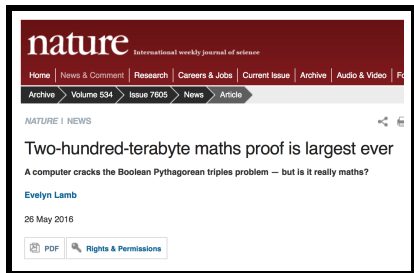
(symbolic AI)



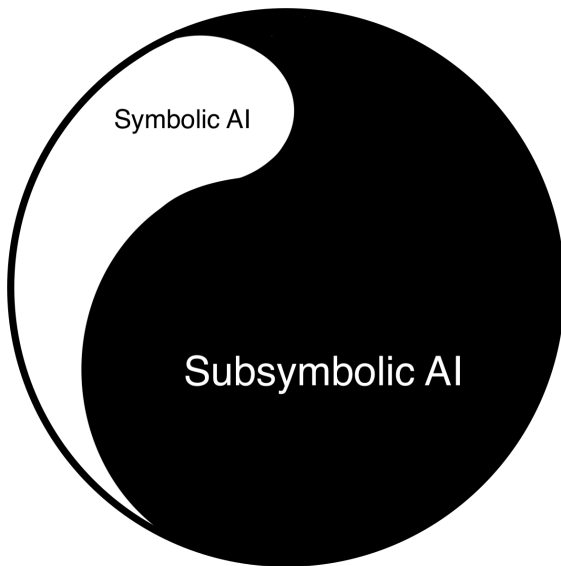
Examples

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

(subsymbolic KI)



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

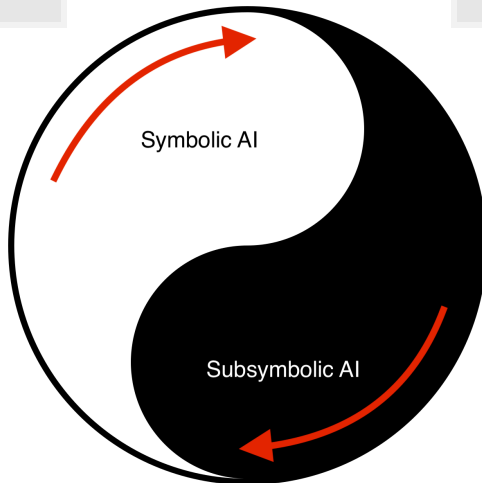


While many sharp minds are already rethinking!

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

Causalities
Abstraction
Precise Reasoning
Explain-/Verifiability
...

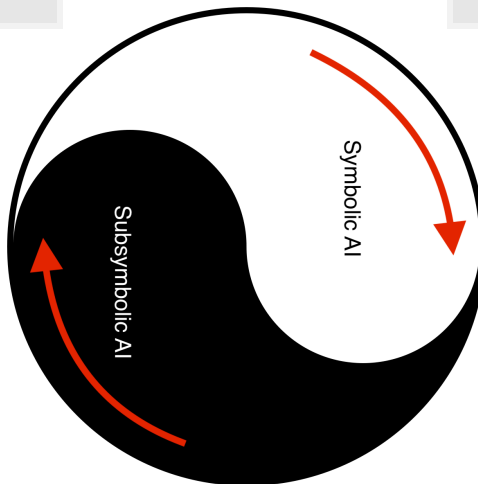
Correlations
Patterns
Robustness
Learning
...



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

Causalities
Abstraction
Precise Reasoning
Explain-/Verifiability
...

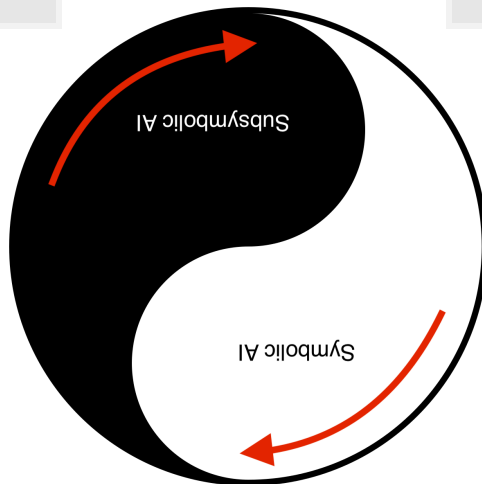
Correlations
Patterns
Robustness
Learning
...



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

Causalities
Abstraction
Precise Reasoning
Explain-/Verifiability
...

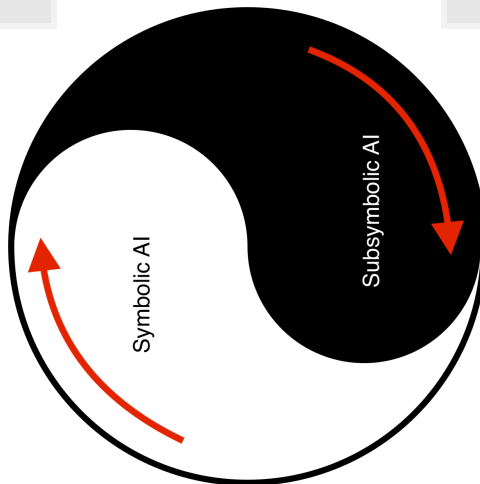
Correlations
Patterns
Robustness
Learning
...



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

Causalities
Abstraction
Precise Reasoning
Explain-/Verifiability
...

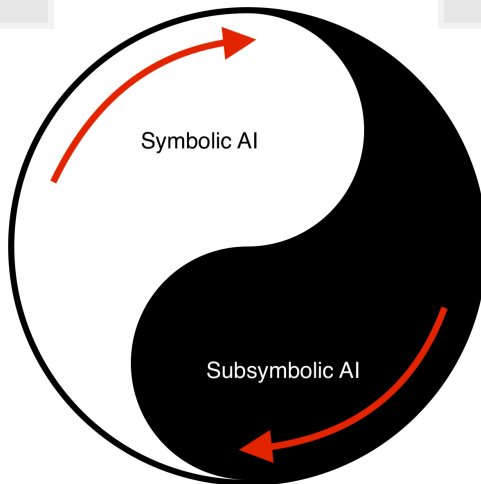
Correlations
Patterns
Robustness
Learning
...



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?



English title: "Rebel Without a Cause" (1955)

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?
- ▶ How can 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?
- ▶ How can 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?
- ▶ How can 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?
- ▶ How can we trust such systems?

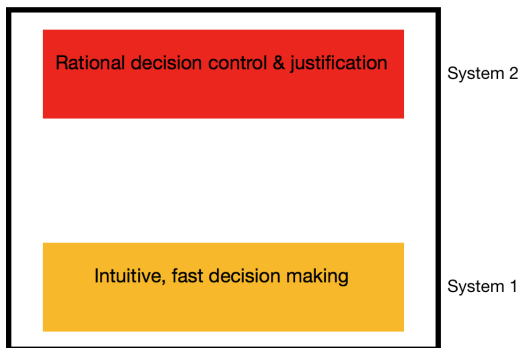
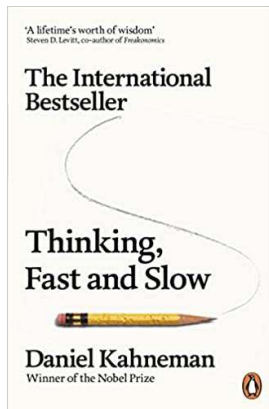
How to create trust?



English title: "Rebel Without a Cause"

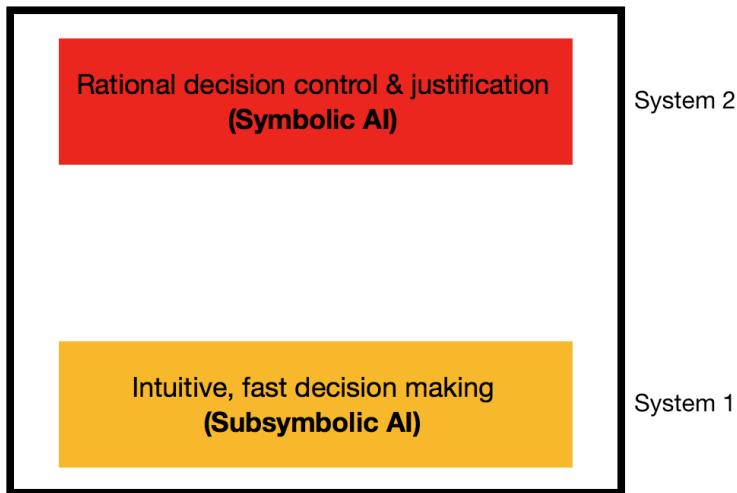
- ▶ Do our current AI systems know what they are doing?
- ▶ Do we know what we are doing when we entrust such AI systems with increasingly critical decisions?
- ▶ Is normative directionlessness and unpredictability the core character of future AI systems?
- ▶ How can we trust such systems?

How to create trust? How to control?

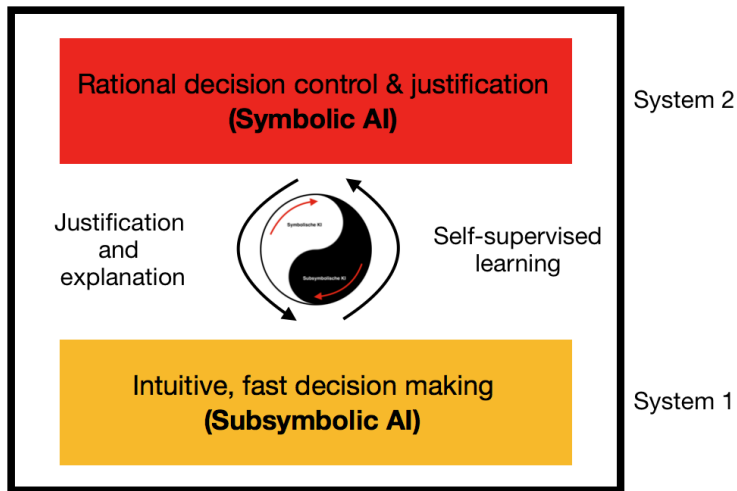


See e.g. also: **J. Haidt, The Emotional Dog and its Rational Tail, 2001**

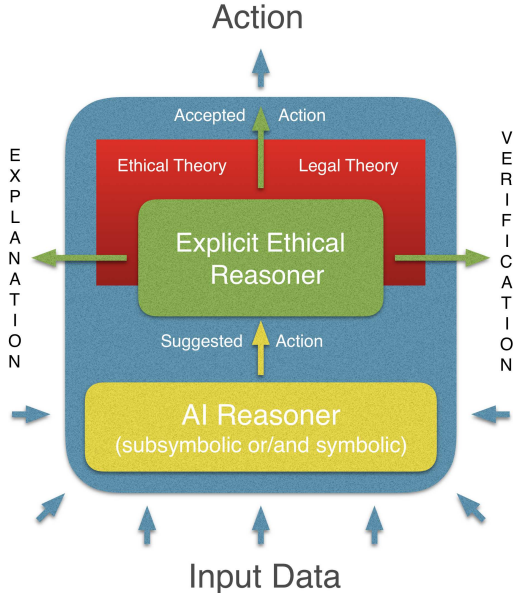
How to create trust? How to control?



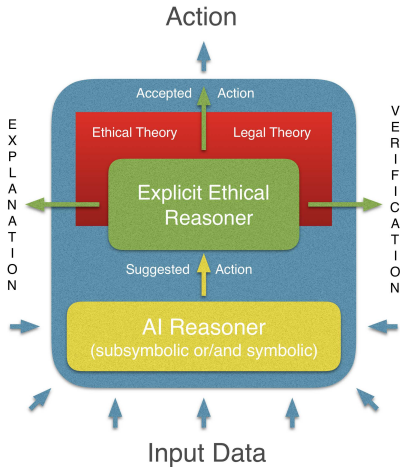
How to create trust? How to control?



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 Benz Müller  , Bertram Lomfeld

[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

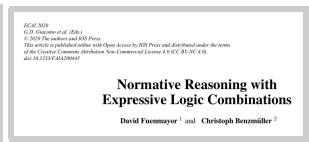
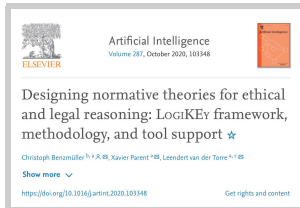
Artificial “Social Reasoning Model”

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

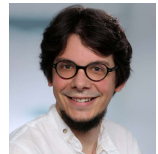
Important:

- ▶ 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



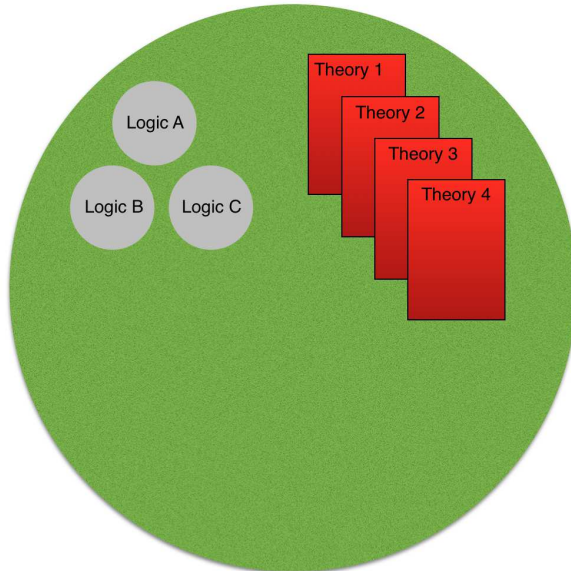
(some recent papers with colleagues from U Luxembourg)



Normative Reasoning Experimentation Framework

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

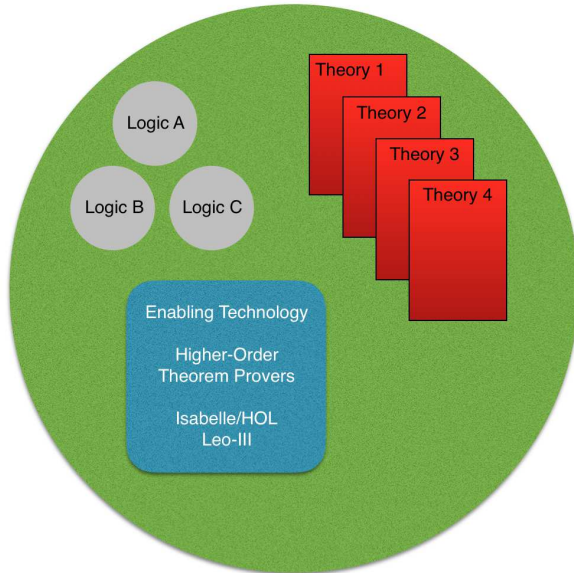
[AIJ (2020) vol. 287], [Data in Brief (2020) vol. 33], www.logikey.org



Normative Reasoning Experimentation Framework

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

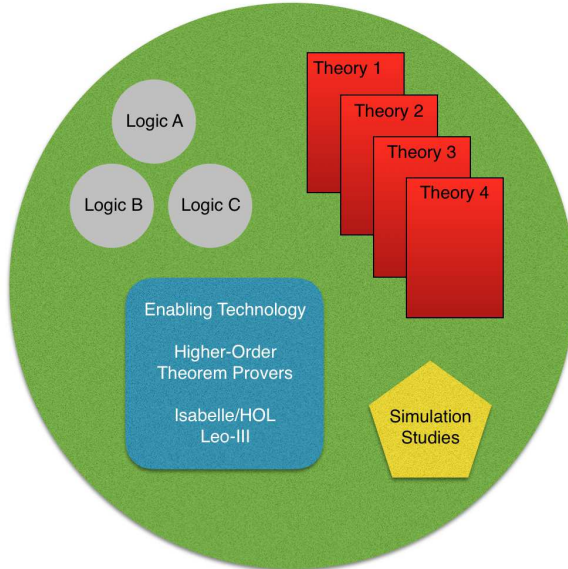
[AIJ (2020) vol. 287], [Data in Brief (2020) vol. 33], www.logikey.org



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



Universal (Meta-)Logical Reasoning in HOL

[SciCompProgr (2019) vol. 172]



How to Tame the Logic Zoo?

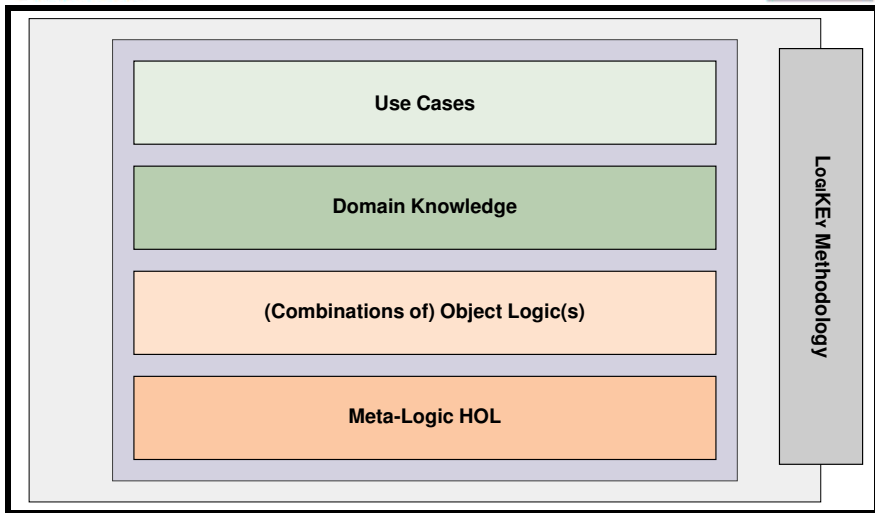
Designing normative theories for ethical and legal reasoning: LOGiKEY framework, methodology, and tool support ☆

Christoph Benzmüller ^{b, a}  , Xavier Parent ^a , Leendert van der Torre ^{a, c} 

Show more 

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

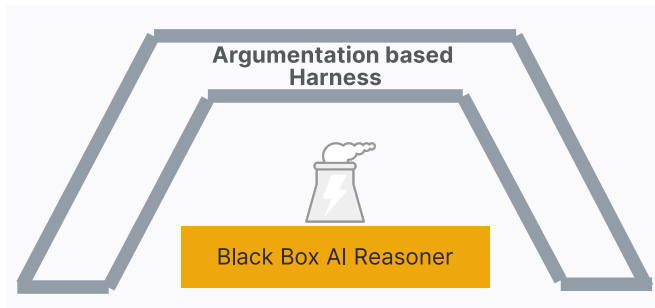
[Get rights and content](#)



Conclusion – How to create trust? How to control?

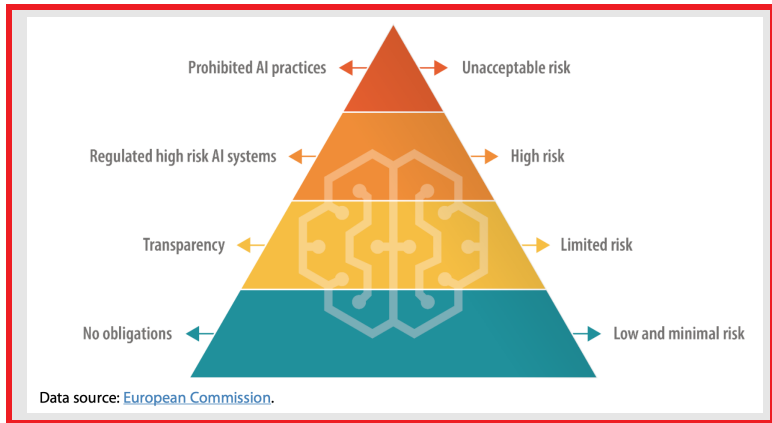
Can post-hoc normative reasoning competencies prevent AI systems from going rogue?

- ▶ 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, justify and communicate 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



Conclusion – How to create trust? How to control?

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



Black Box AI Reasoner