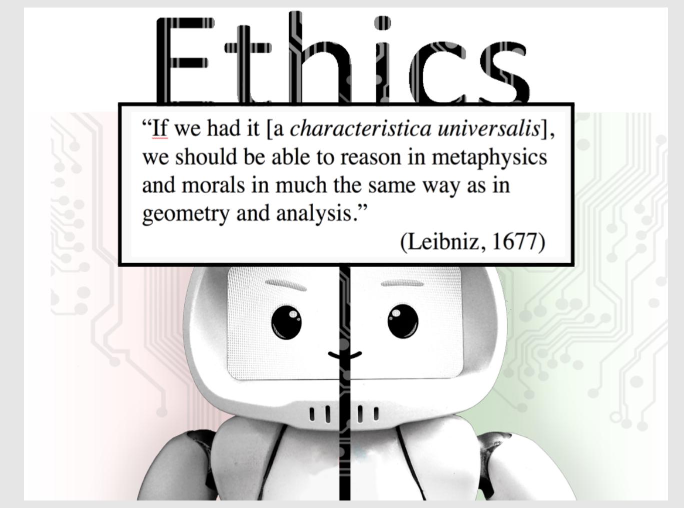
Artificial Intelligence—Definition and Ethico-Legal Aspects

Christoph Benzmüller

Freie Universität Berlin | University of Luxemburg



> "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)

KI 2019

My Interest in AI

Motivation—Siekmann's AI Lecture

"We experience a new branch in evolution, a new species is rising, founded on different physical/biological principles, that will surpass human intelligence in all aspects, and it will be able reproduce itself."

(see also his Duisburg 1985 lecture)

Immediate Reaction:

No way? But how can I refute him?

To study this question further I specialised in a particularly hard area:

Automated reasoning in mathematics, metaphysics and machine ethics

Weak AI, but in a challenge area:

Intuition and creativitity in mathematical proof and theory exploration

A: How to define AI?

Other Definition Attempts

[Alan Turing]

Passing the Turing Test

[Marvin Minsky]

“AI is the science of making machines do things that would require intelligence if done by men.”

[Gottfredson, 1997]

“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.”

A: How to define AI?

Other Definition Attempts

[Joanna Bryson]

“Intelligence is the ability to do the right thing at the right time given a dynamic environment (. . .)”

[Jürgen Schmidhuber]

“A system is intelligent with respect to the standards of a given society with limited physical resources if it can quickly solve or learn to solve a wide variety [of] problems considered relevant and challenging by this society.” –

[Hofstadter and Sander, 2013]

To my mind, intelligence (whether we're speaking of a human or a machine) is the ability to put one's finger on the essence of situations that one faces, and to do so reasonably rapidly. Or a bit more verbosely, ...“intelligence is the art of rapid and reliable gist-finding, crux-spotting, bull's-eye hitting, nub-striking, essence-pinpointing. It is the art of, when one is facing a new situation, swiftly and surely homing in on an insightful precedent (or family of precedents) stored in the recesses of one's memory. That, no more and no less, is what it means to isolate the crux of a new situation. And this is nothing but the ability to find close analogues, which is to say, the ability to come up with strong and useful analogies.

A: How to define AI?

Context dependent notion

Relative use: e.g., one animal is more intelligent than another one
Absolute use: comparison against mental capabilities of humans

Def.: Artificial Intelligence (Benzmüller, March 2019)

Science of computational technologies being developed to achieve and explain *intelligent* behaviour in machines.

Def.: Intelligence (Benzmüller, March 2019)

A collection of mental capabilities that enable an entity

1. to solve (or learn to solve) hard problems, —solve problems—
2. to successfully act in known, unknown and dynamic environments (requires perception, planning, agency, etc.), —master the unknown—
3. to reason abstractly, intuitively, rationally, while avoiding inconsistencies and self-contradiction, —rational&abstract—
4. to reflect upon itself and to adjust its reasoning with upper goals and norms, and —self-reflective—
5. to interact socially with other entities and to align own values and norms with those of a society for a greater good. —be social—

A: How to define AI?

Def.: Intelligence (Benzmüller, March 2019)

A collection of mental capabilities that enable an entity

1. to solve (or learn to solve) hard problems, —solve problems—
2. to successfully act in known, unknown and dynamic environments (requires perception, planning, agency, etc.), —master the unknown—
3. to reason abstractly, intuitively, rationally, while avoiding inconsistencies and self-contradiction, —rational&abstract—
4. to reflect upon itself and to adjust its reasoning with upper goals and norms, and —self-reflective—
5. to interact socially with other entities and to align own values and norms with those of a society for a greater good. —interact socially—

Remarks

- Chatbot Tay (later replaced by Zo) would not qualify as very intelligent
- 3.-5. enable IAS to identify and refuse malicious/nonsense training data.
- Example: I refused in my early economics lectures to accept the idea that "air" is a good example for a "free good" (societal costs is polution)
- "Information" should also be handled with care: no free good

B: Why is it that research and development in AI is booming?

Or does that only seem so?

- Steady development of AI technology in all subfields for decades
  - Germany: Industry 4.0 and the DFKI
  - Internationally: SAT solving technology now applied widely in Industry
- Success story in machine learning: AI solutions with little effort
- Showcase projects by large corporations (e.g. IBM):

  2011 Jeopardy won by Watson — broadcasted on TV on large scale

  2016 Go won by AlphaGo, later by AlphaZero
- AI technology has arrived in everyday life: Siri, Alexa, etc.
- Excellent topic for media:

  Science Fiction becoming Reality

  Machine Learning easy to Depict
- Statements by Prominent Figures:
  - Gates, Hawkins, Musk, Bostrom: Emerging Superintelligence
  - Putin: Nation that leads in AI ‘will be the ruler of the world’
  - Investments and dynamics in China
  - Special AI departments/units now being installed everywhere
  - Start-ups; my students disappeared
  - Industry is hiring

B: AI Success Story in Germany

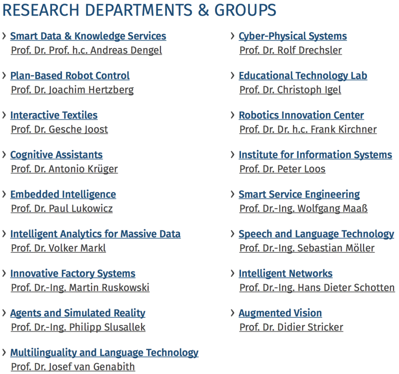
Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

Standort Kaiserslautern *Standort Saarbrücken* *Standort Bremen* *Projektbüro Berlin*

Profile

- success story since 1988
- Kaiserslautern, Saarbrücken, Bremen Berlin, Osnabrück, St. Wendel
- Budget 2017: 45,9 Mill. €
- 1000 Employees, 60 Nations
- created 130 Profs, 90 Spin-offs

Inspired e.g. Industry 4.0 Topic



RESEARCH DEPARTMENTS & GROUPS

› Smart Data & Knowledge Services
Prof. Dr. Prof. h.c. Andreas Dengel

› Plan-Based Robot Control
Prof. Dr. Joachim Hertzberg

› Interactive Textiles
Prof. Dr. Gesche Joost

› Cognitive Assistants
Prof. Dr. Antonio Krüger

› Embedded Intelligence
Prof. Dr. Paul Lukowicz

› Intelligent Analytics for Massive Data
Prof. Dr. Volker Markl

› Innovative Factory Systems
Prof. Dr.-Ing. Martin Ruskowski

› Agents and Simulated Reality
Prof. Dr.-Ing. Philipp Slusallek

› Multilinguality and Language Technology
Prof. Dr. Josef van Genabith

› Cyber-Physical Systems
Prof. Dr. Rolf Drechsler

› Educational Technology Lab
Prof. Dr. Christoph Igel

› Robotics Innovation Center
Prof. Dr. Dr. h.c. Frank Kirchner

› Institute for Information Systems
Prof. Dr. Peter Loos

› Smart Service Engineering
Prof. Dr.-Ing. Wolfgang Maaß

› Speech and Language Technology
Prof. Dr.-Ing. Sebastian Möller

› Intelligent Networks
Prof. Dr.-Ing. Hans Dieter Schotten

› Augmented Vision
Prof. Dr. Didier Stricker

C: What are the Challenges Ahead?—Personal Opinion

1. Assess and Monitor: Impact of AI on Humanity and Societies
2. Educate, Educate, Educate (interdisciplinary & whole society)
3. Cern for AI, Compete for best Talents
4. Find Optimal Balance: Opportunities vs. Risks & Costs
5. Set up Rules for Thrustworthy AI made in Europe
6. Clear positioning to Area of Autonomous (Cyber-)Warfare Systems
7. Adjust Social Systems
8. Individual Plans for Individual Fields
9. Next Big Thing: Integration of Symbolic & Subsymbolic AI (?)

C: What are the Challenges Ahead?

1. Assess and Monitor: Impact of AI on Humanity and Societies

- Clear: there will be an impact (already now)
- Unclear: how much, who affected, where, when?
- Policy makers must be ahead; *"develop, then regulate"* not a good scheme

2. Educate, Educate, Educate: Interdisciplinary!!!

- new generation of policymakers and academics needed
- informed society needed (prevent thread to democracy)
- Example: PhD programme with FU Kennedy-Institute (Gienow-Hecht)
- rethink the antiquated structures in academia

3. Cern for AI, Compete for best talents

- Cutting edge, shared research infrastructure for AI
- See the objective of the CLAIRE initiative

C: What are the Challenges Ahead?

4. Find Optimal Balance: Opportunities vs. Risks & Costs

Opportunities

- increased productivity, economic growth
- transport, healthcare, climate, energy, finance, government, education, etc.

Risk & Costs

- privacy and security, vulnerability of democratic systems
- concentration: power, knowledge
- social divide: wealth, access to information & technology

5. Set up Rules for Thrustworthy AI made in Europe

Industry will benefit from a clarity (liability, responsibility, design principles)

6. Clear positioning to Area of Autonomous (Cyber-)Warfare Systems

Discussion under 2. only makes sense if 3. is not excluded

C: What are the Challenges Ahead?

7. Adjust Social Systems

- How can social divide be prevented and stopped?
- High dynamics of change
- Taxation of AI technology? Basic Income Scheme?
- Rethink: full employment and personal identification with job
- Prepare for a society with more individual free time

8. Individual Plans for Individual Fields

- Autonomous (Cyber-)Warfare vs. Autonomous Cars
- Financial trading, social scoring, personal care, etc.
- How much autonomous agency do we want to allow in which field

Next big thing: Integration of Symbolic & Subsymbolic AI

- Leading corporations and companies are already on this topic

D: When will there be self-driving cars (in Berlin)?

- Continuous process
- Will eventually start soon after regulation issues have been solved
- Tests phases?—restricted areas at restricted times
- Car-sharing, e.g. optimal car relocation at night
- Public transport—reserved bus

E: Are there other AI applications to be expected to make cities sustainable??

- Intelligent integrated transportation system
- Intelligent energy consumption and production
- Intelligent governance & administration
- Intelligent delivery systems for groceries and other household goods

F: How far are we with developing legal frameworks for AI?

EU’s General Data Protection Regulation (GDPR)

- Art. 17—Right to be forgotten,
- Recital 71—Right for explanation

EU Parliament (Resol., 1/2019) Autonomous driving in European transport

- current regulatory framework will presumably not be sufficient
- ethical aspects need to be addressed and resolved by the legislator before these vehicles can be fully accepted
- automated vehicles need to undergo assessment of ethical aspects

HLEG (Expert Group): Ethics Guidelines

- ensure an “ethical purpose”
- ensure technical robustness and reliability

AI4People (Global Public Forum): Ethical Framework for “Good AI Society”

- define ethical framework
- recommendations how to implement such a framework

F: EU Parliament—Resolution of 15. Jan 2019

Autonomous driving in European transport (2018/2089(INI))

The European Parliament,

20. Notes that the **existing liability rules**, such as . . . , **were not developed to deal with the challenges posed by the use of autonomous vehicles** and stresses that there is growing evidence that the current regulatory framework, especially as regards liability, insurance, registration and protection of personal data, **will no longer be sufficient** or adequate when faced with the new risks emerging from increasing vehicle automation, connectivity and complexity;

21. . . . calls, therefore, on the Commission to . . . introduce, if necessary, **new rules on the basis of which responsibility and liability are allocated**; calls also on the Commission to assess and monitor the possibility of introducing additional EU instruments to keep pace with developments in AI;

35. Calls on the Commission to lay down clear ethical guidelines for AI;

37. Stresses that **ethical aspects of self-driving vehicles need to be addressed and resolved by the legislator before these vehicles can be fully accepted** and made available in traffic situations; emphasises, therefore, that automated vehicles need to **undergo a prior assessment** to address these ethical aspects;

F: HLEG—High-Level Expert Group on AI

Ethics Guidelines for Thrustworthy AI

Summary of European Position: Thrustworthy AI made in Europe

- maximise benefits of AI while minimising the risks
- human-centric approach to AI
- **Thrustworthy AI** is the north star

Component 1: It should respect fundamental rights, applicable regulation and core principles and values, ensuring an **"ethical purpose"**;

Component 2: It should be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm.

F: HLEG—High-Level Expert Group on Artificial Intelligence

Ethics Guidelines for Thrustworthy AI

Notion of Trust

- in technology, through the way it is built
- in rules, laws, norms that govern AI
- in business and public governance models of AI services, products, and manufacturers

Def. Ethics

What is Good? What is the right thing to do? What is good life?

Ethical Principles

- Beneficence: *Do Good*
- Non-Maleficence: *Do no Harm*
- Autonomy: *Preserve Human Agency*
- Justice: *Be Fair*
- Explicability: *Operate Transparently*

F: CLAIRE—Confederation of Laboratories for AI Research in Europe

claire-ai.org

Initiative by the European AI community that seeks to strengthen European excellence in AI research and innovation; proposes the establishment of a pan-European Confederation of Laboratories for AI Research in Europe that achieves “brand recognition” similar to CERN

It is not only Europe that needs AI made in Europe.

CLAIRE will

- focus on trustworthy AI that augments human intelligence rather than replacing it, and that thus benefits the people of Europe;
- work for a major increase in funding towards existing scientific strengths in AI, novel research opportunities and key European interests;
- work with key stakeholders to find mechanisms for citizen engagement, industry and public sector collaboration and innovation-driven startup and scale-up;
- in this way define and address challenges in various sectors and across a wide range of applications, including health, manufacturing, transportation, scientific research, financial services and entertainment.

Supporters: 1616 AI experts and 668 supporters in industry

F: AI4People—Ethical Framework for "Good AI Society"

http://www.eismd.eu/ai4people/

AI4People (launched November 2017) is the first global forum in Europe on the social impact of AI; goal is to create a common public space for laying out the founding principles, policies and practices on which to build a "good AI society".

AI4People Goals

- identify core values that should inform an ethical framework supporting a preferable development of AI,
- designing a European ethical framework for "good AI society", and
- recommending actionable measures for successful implementation of the framework.

Download: Ethical Framework for "Good AI Society"

F: AI4People—Ethical Framework for "Good AI Society"

http://www.eismd.eu/ai4people/

- Development of Legal and Ethical Compliance
  - Legal compliance: least that must be done
    (GDPR: Art. 17—Right to be forgotten, Recital 71—Right for explanation)
- Ethical Approach to AI
  - alignment of development of AI technology with societal norms and values
  - creates new opportunities and prevents undesired costs
  - early warning sytem on societies adoption of AI technology
- **Question:** Ethical and Legal Rules for Whom?
  - developers of AI technology (Frankenstein)
  - AI technology itself (Frankenstein's monster)
  - **Answer:** both is needed
  - dependency on degree of autonomous agency we affirm to AI technology
- Possible Rules for Developers
  - implementation quality standards
  - means to ensure fairness, transparency, explainability, security, etc.
  - formal verification

F: Own Research Interest—Ethical Governors

- Independent, self-reflectory reasoning normative competency
- Applies formal methods in normative reasoning
- Layer of Machine Learning vs. Layer of Trust, Values and Norms
- Challenges
  - Paradoxes in normative reasoning
  - Formalisation of normative theories: expressive non-classical logics
  - Reasoning with normative theories: consistency, compliance, entailment
  - Alignment of normative theories and values
  - Simulation studies: Which normative theory works?
  - Efficient and secure deployment in systems

F: Own Work in Machine Ethics

Alan Gewirth's *"Principle of Generic Consistency"* (PGC)

Emendation of the Golden Rule

Principle of treating others as one's self would wish to be treated.

PGC is a closely related upper moral principle:

Any intelligent agent, by virtue of its self-understanding as an agent, is rationally committed to asserting that

(i) it has rights to freedom and well-being, and

(ii) all other agents have those same rights.

PGC has lately been proposed as a means to bound the impact of artificial general intelligence (AGI).

PGC has been formalised and (partially) automated in Isabelle/HOL

F: Asilomar AI Principles—https://futureoflife.org/ai-principles/

Developed in conjunction with the 2017 Asilomar conference

Research Issues

1. Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
2. Research Funding: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as: How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked? How can we grow our prosperity through automation while maintaining people's resources and purpose? How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI? What set of values should AI be aligned with, and what legal and ethical status should it have?
3. Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policy-makers.
4. Research Culture: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
5. Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

F: Asilomar AI Principles—`https://futureoflife.org/ai-principles/`

Developed in conjunction with the 2017 Asilomar conference

Ethics and Values

6 Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

7 Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.

8 Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9 Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10 Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11 Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

F: Asilomar AI Principles—`https://futureoflife.org/ai-principles/`

Developed in conjunction with the 2017 Asilomar conference

Ethics and Values (cont'd)

12 Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13 Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14 Shared Benefit: AI technologies should benefit and empower as many people as possible.

15 Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

16 Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

17 Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

18 AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

F: Asilomar AI Principles—https://futureoflife.org/ai-principles/

Longer-term Issues

19 Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

20 Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

21 Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

22 Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

23 Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.