Praxisseminar SoSe 2016 Technologiefolgeabschätzung für die KI

COLI, Uni Saarland

Von Thierry Declerck

Künstliche Intelligenz (KI)

- **Künstliche Intelligenz** (**KI**, englisch *artificial intelligence*, *AI*) ist ein Teilgebiet der Informatik, welches sich mit der Automatisierung intelligenten Verhaltens befasst. Der Begriff ist insofern nicht eindeutig abgrenzbar, da es bereits an einer genauen Definition von Intelligenz mangelt. Dennoch findet er in Forschung und Entwicklung Anwendung.
- Im Allgemeinen bezeichnet „künstliche Intelligenz“ oder „KI“ den Versuch, eine menschenähnliche Intelligenz nachzubilden, d. h., einen Computer zu bauen oder so zu programmieren, dass dieser eigenständig Probleme bearbeiten kann. Oftmals wird damit aber auch eine effektvoll nachgeahmte, vorgetäuschte Intelligenz bezeichnet, insbesondere bei Computerspielen, die durch meist einfache Algorithmen ein intelligentes Verhalten simulieren soll. (Wikipedia: http://de.wikipedia.org/wiki/K%C3%BCnstliche_Intelligenz)

Technologiefolgenabschätzung

- Das Forschungsgebiet der **Technikfolgenabschätzung** (kurz TA, auch: *Technologiefolgenabschätzung* oder *Technikbewertung*) ist ein Teilgebiet der Technikphilosophie und -soziologie. Es entstand in den 1960er Jahren in den USA und verbreitete sich von den 1970er Jahren an in Europa.
- Die Technikfolgenabschätzung befasst sich mit der Beobachtung und Analyse von Trends in Wissenschaft und Technik und den damit zusammenhängenden gesellschaftlichen Entwicklungen, insbesondere der Abschätzung der Chancen und Risiken.
- Zudem soll die Technikfolgenabschätzung politische Handlungsempfehlungen oder Richtlinien für die Vermeidung von Risiken und die verbesserte Nutzung der Chancen geben (siehe auch Gefährdung). Damit stellt sie eine konzeptionelle Erweiterung der klassischen Entscheidungstheorie dar.

(http://de.wikipedia.org/wiki/Technikfolgenabsch%C3%A4tzung)

Technologiekritik

- **Critique of technology** is an analysis of the negative impacts of technologies. It is argued that, in all advanced industrial societies (not necessarily only capitalist ones), technology becomes a means of domination, control and exploitation, or more generally something which threatens the survival of humanity. (Wikipedia)
- Und persönliche Einstellungen: für wen/was arbeite ich? Offene und transparente Forschung, oder geheimgehalten (Firmenstrategien, Militär, Geheimdienst, …)

Kritikpunkte an der KI

- Moralische und Religiöse Fragestellungen zum Thema „nicht menschliche Intelligenz“. Darf man so was kreieren wollen?
- Entscheidungen werden mehr und mehr Maschinen überlassen, zum Bsp. im autonomen Fahren, Finanzbereich oder beim Militär
  - Frage der juristischen Verantwortung: Börsenkrach, Tötung von Zivilisten bei automatisierten Waffensystemen, …
- Aber vielleicht ist die Technologie nicht gut oder böse, sondern nur diejenigen, die sie anwenden?

Aber Faszination des Menschen für das autonomen Schaffen von „Leben“

- Golem (aber endet schlecht für Golem, wenn er zum Monster mutiert)
  - In Jewish folklore, a **golem** (/ˈgoʊləm/ ***GOH****-ləm*; Hebrew: גולם) is an animated anthropomorphic being, magically created entirely from inanimate matter (Wikipedia)
- Frankenstein (aber es entwickelt sich schlecht für die „Schöpfung“ von Frankenstein, und manchen beteiligten)
- Aber auch in den Bio-Wissenschaften: Cloning, Künstliche Befruchtung, Maschine-Assistierten Reproduktion, etc.
  - Geht das gut?

Gefahren durch Maschinen

- 2001 Space Odyssee
  - HAL übernimmt Kommando und tötet die mitreisenden Menschen (bis auf einen, der dann HAL „abschaltet“). Video!
  - Interessanter Aspekt: Geschichte der Intelligenz ist Geschichte der Gewalt/des Tötens.
- Verletzung der Robotik-Gesetzen (Isaac Asimov)
  - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  - A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
  - A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
- In 2001 protegiert HAL 9000 die Mission des Raumschiffes, und betrachtet die mitreisenden als Gefahr für die Mission.

Gefahren durch die KI

- Joseph Weizenbaum
  - Buch „***Computer Power and Human Reason: From Judgment To Calculation***“ (1976), und andere Werke
- Hubert Dreyfuss
  - Buch „ ***What Computers Can't Do*** (1972; 1979; 1992), und andere Werke
- Stephen Hawking „ The development of full artificial intelligence could spell the end of the human race”
- A related Open Letter by Future of Life Institute (LFI) (2014) signed by many scientists, but also with the suggestions of some guidelines to keep AI beneficial to humans.

Weizenbaum und „Eliza“

- **ELIZA** is a computer program and an early example of primitive natural language processing. ELIZA operated by processing users' responses to *scripts*, the most famous of which was **DOCTOR**, a simulation of a psychotherapist. Using almost no information about human thought or emotion, DOCTOR sometimes provided a startlingly human-like interaction. ELIZA was written at MIT by Joseph Weizenbaum between 1964 and 1966.
- When the "patient" exceeded the very small knowledge base, DOCTOR might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?" A possible response to "My mother hates me" would be "Who else in your family hates you?" ELIZA was implemented using simple pattern matching techniques, but was taken seriously by several of its users, even after Weizenbaum explained to them how it worked. It was one of the first chatterbots.
- (wikipedia:

Joseph Weizenbaum's influential 1976 book ***Computer Power and Human Reason: From Judgment To Calculation***

- displays his ambivalence towards computer technology and lays out his case: while artificial intelligence may be possible, we should never allow computers to make important decisions because computers will always lack human qualities such as compassion and wisdom.
- Weizenbaum makes the crucial distinction between deciding and choosing. Deciding is a computational activity, something that can ultimately be programmed. It is the capacity to choose that ultimately makes us human.
- Choice, however, is the product of judgment, not calculation. Comprehensive human judgment is able to include non-mathematical factors such as emotions. Judgment can compare apples and oranges, and can do so without quantifying each fruit type and then reductively quantifying each to factors necessary for mathematical comparison.
- (wikipedia)

Hubert Dreyfus

- Dreyfus argued that human intelligence and expertise depend primarily on unconscious instincts rather than conscious symbolic manipulation, and that these unconscious skills could never be captured in formal rules. His critique was based on the insights of modern continental philosophers such as Merleau-Ponty and Heidegger, and was directed at the first wave of AI research which used high level formal symbols to represent reality and tried to reduce intelligence to symbol manipulation.
- When Dreyfus' ideas were first introduced in the mid-1960s, they were met with ridicule and outright hostility. By the 1980s, however, many of his perspectives were rediscovered by researchers working in robotics and the new field of connectionism—approaches now called "sub-symbolic" because they eschew early AI research's emphasis on high level symbols. Historian and AI researcher Daniel Crevier writes: "time has proven the accuracy and perceptiveness of some of Dreyfus's comments." Dreyfus said in 2007 "I figure I won and it's over—they've given up."
- (Wikipedia)

The four assumptions critized by Dreyfus. They all ignore „intuition“

- The biological assumption: *The brain processes information in discrete operations by way of some biological equivalent of on/off switches.*
- The psychological assumption: *The mind can be viewed as a device operating on bits of information according to formal rules.*
- The epistemological assumption: *All knowledge can be formalized.*
- The ontological assumption: *The world consists of independent facts that can be represented by independent symbols*

Die große Frage:

- Funktioniert der menschliche Geist, wie ein Computer, und kann ein Computer den menschlichen Geist simulieren/modellieren

Wettkampf Computer vs Geist

- Computer können jetzt (aber erst jetzt, aber nicht wie in den 60er gedacht für sehr nah prognostiziert)
  - Deep Blue
- Computer können Menschen bei Quizspielen schlagen
  - Das Q&A Watson System von IBM
- Aber wohl mehr sehr starke Berechnungsfähigkeiten, und eher weniger Denken bei diesen neueren Algorithmen.

Stephan Hawking on BBC:
artificial intelligence could end mankind

- Prof Hawking says the primitive forms of artificial intelligence developed so far have already proved very useful, but he fears the consequences of creating something that can match or surpass humans.

- "It would take off on its own, and re-design itself at an ever increasing rate," he said.

- "Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded."

Can Artificial Intelligence cause Human Extinction? (assumption by Stephen Hawking) Tale by Jordi Torras

- Step 1 **A.I. is born:** Technologies like Natural Language Processing, Artificial Neural Networks, and Computer Vision arise. They can make computers speak, understand and solve complex problems, drive cars, fly planes, sail boats and execute a variety of activities that only highly trained humans had been able to perform before.
- Step 2 **Evolution:** Genetic Algorithms, which only allow the most efficient applications to pass on their "cyber-genes" to the next generation, improve the software/hardware combinations by using the same principles as natural selection. Improving AI technologies by only allowing the best ones to exist is not very different from what humans have previously done with dogs, cows, horses and so many domestic animals over thousands of years.
- Step 3 **Military use:** In order to prevent human losses, intelligent machines are sent to war instead of humans. These wars take place in dangerous or inaccessible terrain, and occur too far away from where the armies can control and monitor the troops, so more and more autonomous machines are required. For simplicity, let's call these intelligent war machines "w-Droids", and let's call that hypothetical war "The War".

- Step 4 **Self-replication:** As The War develops and the battles get longer and harder to win; producing new droids locally, close to where battles take place, is a great advantage. Also, in order to improve efficiency and the chances to win The War, the w-Droids that survive combats and battles are used to produce the next generation of the w-Droids using the "genetic algorithms". Over time, factories used to produce w-Droids are also managed and controlled by w-Droids.
- Step 5 **The survival of the fittest:** Soon, some specific "cyber-genes" start to spread and are very common in w-Droids populations. For example, genes that make w-Droids tend to protect themselves, cyber-genes that make them protect other w-Droids that they recognize as having their own cyber-gens (let's call it, protecting their "offspring"), would be a great advantage for them.
- Step 6 **Isolation:** As The War gets even more cruel, human troops are gradually sent home. Slowly but firmly, journalists are also replaced by w-Droids, since taking images from battles becomes more and more dangerous. After a few years of The War, w-Droids become isolated, and the only information the general public can obtain from inside The War is sent by w-Droids themselves.

- Step 7 **The discovery:** One day, a terrible fact is discovered: the war had been over long ago, but w-Droids kept sending digitally-created videos of war scenes where the enemies are crunched by w-Droids, and grateful, smiling children are saved by the w-Droids. At that point in time, no human exactly knows what the w-Droids are doing, except for a few remaining humans that still survive in the country where The War is happening.
- Step 8 **A new Nation:** Governments urge retreat. W-Droids are told to leave the land and retreat, but something terrible happens: W-Droids refuse to come back. No human is now left in the former "enemy territory", and w-Droids declare the first "w-Droid Nation", which has formidable military power and probably includes nuclear weapons. At that point, if the w-Droids decide to keep fighting, the end of human existence is inevitable and might take place only a few decades later

Open Letter from FLI (Future Life Institute)

- Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents - systems that perceive and act in some environment. In this context, "intelligence" is related to statistical and economic notions of rationality - colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.
- As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.
- The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008-09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do. The attached research priorities document gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law and philosophy to computer security, formal methods and, of course, various branches of AI itself.
- In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.

FLI – Research Priorities

- The Research Priorities Document by FLI

KI und Arbeitsplätze

- Jede neue Technologie hat Einfluss auf die Arbeitswelt und auch auf die Gesellschaft/Kultur
  - Vor vielen Jahren, Schriftsprache kritisiert, weil das Denken verringernd
  - Vor vielen Jahren das Druckbuch „killt“ einige Jobs (Bibelkopierer, Clercks), aber demokratisiert und kreiert neue Arten von Jobs.
  - Mit Computertechnologien, wohl ähnlich, aber
- “Unchecked, automation will destroy many skilled jobs and push people into insecure, unskilled work. The winners will be those who learn how to code” (Jim Chalmers and Tim Watts)
- It’s a result with a worrying social aspect, with the possibility of the workforce being divided between “those who are good at working with intelligent machines, and those who are replaced by them” – the grim world that economist Tyler Cowen has described.
  - Dauerlernen ist heutzutage erforderlich, um mit der Technologie Entwicklung mithalten zu können

- Danke für die Diskussion ☺