Praxisseminar SoSe 2017 Technologiefolgeabschätzung für die KI

COLI, Uni Saarland

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 -- acces 13.05.2017)

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 – Access 13.05.2017)

Eine erste “Technologie Kritik” bereits bei Plato

http://www.gradesaver.com/phaedrus/study-guide/the-technology-of-writing

- “But new technology often breeds suspicion. Much like contemporary anxiety over the increasingly “cold” modes of electronic communication (e.g., email in the place of postal mail), Thamus’s dislike of writing in the myth of Theuth is often taken to reflect Plato’s suspicion of writing as a new technology. Jacques Derrida’s famous essay “Plato’s Pharmacy,” for example, treats Plato’s ostensible preference of speech over writing as one of its subjects.
- The changes induced by new technology, however, often catch us unawares. In the influential study Orality and Literacy, Walter J. Ong first glosses Plato’s criticism of writing. According to Ong, the Phaedrus and so-called Seventh Letter raise four main points: as opposed to speech, writing is inhuman, a thing, a technological product; it weakens the memory of those who rely on it; it cannot respond to new questions; and it cannot defend itself (274-77). Writing is cast essentially as a passive, impersonal product that serves as a poor substitute for speech. For Plato to make his objections strongly and effectively, however, he himself chose to use writing (albeit in dialogue form and using characters other than himself who are speaking). This allowed him to concretize and develop his ideas in ways that were perhaps unavailable through direct speech. Moreover, we would otherwise not have his objections passed down to us in the way he intended them.
- Consequently, as Ong points out (based on Eric Havelock’s study Preface to Plato), the use of writing unwittingly turned Plato against the former oral tradition:
- Plato’s entire epistemology was unwittingly a programmed rejection of the old oral, mobile, warm, personally interactive lifeworld of oral culture . . . Platonic form was form conceived of by analogy with visible form. The Platonic ideas are voiceless, immobile, devoid of all warmth, not interactive but isolated, not part of the human lifeworld at all but utterly above and beyond it. (80)

Eine erste “Technologie Kritik” bereits bei Plato (2)

http://www.gradesaver.com/phaedrus/study-guide/the-technology-of-writing

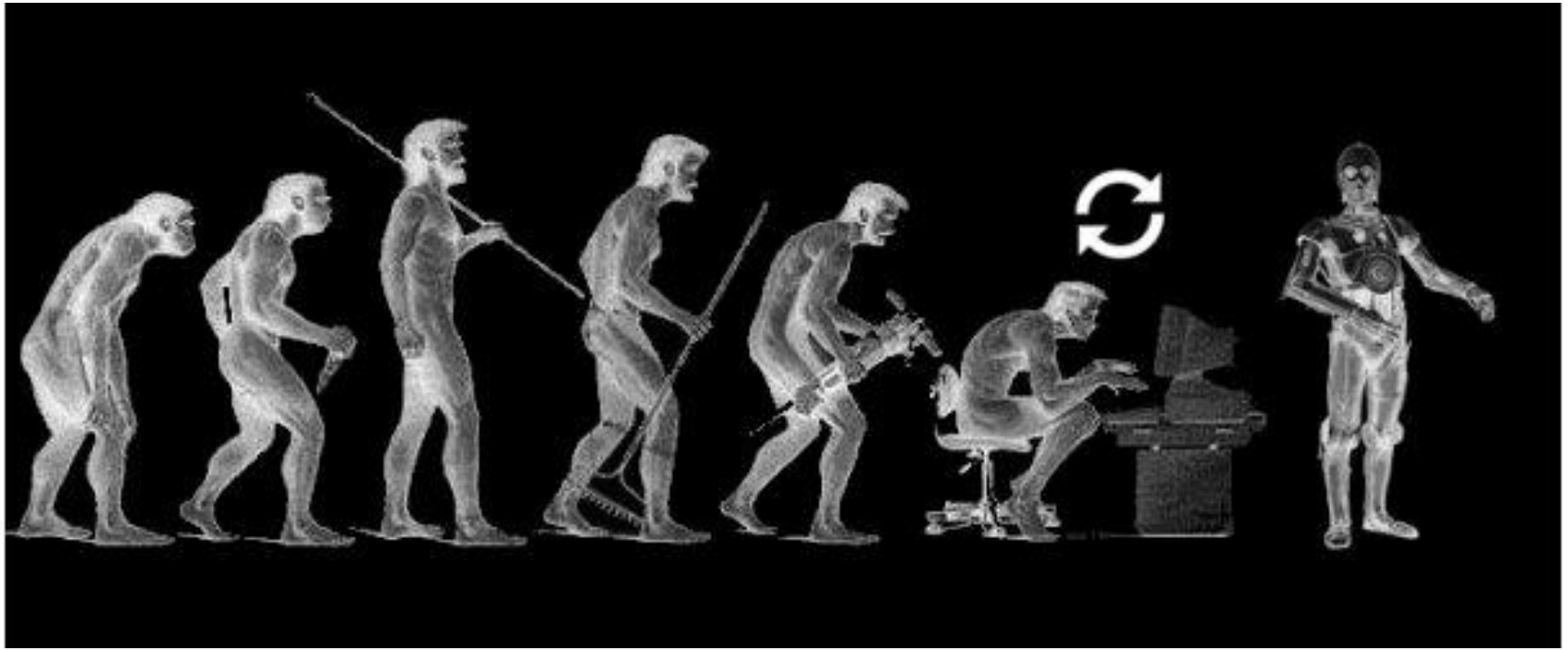
- By turning to writing, Plato was inadvertently influenced by the very paradigm of literacy he opposed—a paradigm based on seeing rather than hearing. “The term idea, form,” Ong reminds us, “is visually based, coming from the same root as the Latin *video*, to see” (80). The Platonic Idea resembles writing in that it is absolute and autonomous. Like writing, it can be perceived and talked about; unlike speech, it has no immediate presence on a human level.
- Ultimately, history has resolved the disagreement between Theuth and Thamus in favor of the former. As Ong shows in Orality and Literacy, writing has restructured human consciousness in a way that has increased both wisdom and cultural memory.
- Such historical arguments remain entirely relevant. With regard to new modes of communication, readers today may find themselves in a position akin to that of Plato (through Socrates, a generation his senior) in the Phaedrus. Essentially every argument that Plato makes against writing can be made analogously against the Internet and electronic communications.
- Consider, for instance, what "knowing facts" means in the context of being able to retrieve all sorts of facts from a handheld Internet connection at any time, anywhere. When Plato has Socrates say elsewhere that he knows many things but they are all trivial, perhaps he means things that can be written down, like lists of rhetorical devices. Knowing love or thinking philosophically is something else entirely. What does the future hold for an increasingly technological world, where more and more can be recorded or calculated outside of our minds?

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. (https://en.wikipedia.org/wiki/Critique_of_technology, Access 13.05.2017)
- Und persönliche Einstellungen: für wen/was arbeite ich? Offene und transparente Forschung, oder geheingehalten (Firmenstrategien, Militär, Geheimdienst, …)

KI Vision der menschlichen Evolution ?

http://www.pabloferreiragonzalez.com/2016/11/01/awesome-artificial-intelligence/



Mögliche Folgen der KI: Evolution (2)

https://www.mobilegeeks.de/artikel/arbeitslos-durch-intelligente-maschinen/

Kritikpunkte an die 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 Finanzbereich oder bei Militär
  - Frage der juristischen Verantwortung: Börsenkrach, Tötung von Zivilisten bei automatisierten Waffensystemen, …
- Aber vielleicht ist die Technologie nicht gut oder böse, sondern nur diejenige, die sie anwenden?

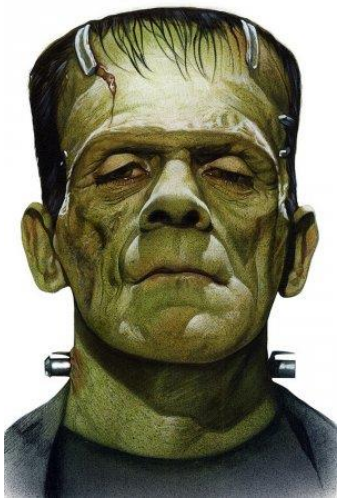
Beispiel einer Computer gestützten fatalen Entscheidung: Iran-Air-Flug 655

- “Der von Iran-Air-Flug 655 als Startpunkt genutzte Flughafen von Bandar Abbas wurde damals sowohl von Zivil- als auch von Militärflugzeugen genutzt.[1] Der Airbus wurde durch eine automatische Anfrage der *Vincennes* beim Transponder der Linienmaschine als Zivilflugzeug erkannt, jedoch identifizierte das Aegis-Kampfsystem der *Vincennes* eine F-14 Tomcat. Die Besatzung der *Vincennes* entschied sich, der Meldung des Aegis-Systems zu glauben.“(https://de.wikipedia.org/wiki/Iran-Air-Flug_655, Access 13.05.2017).
- Das Flugzeug mit 290 Insassen wurde abgeschossen.

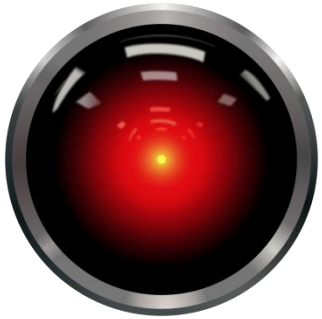
Aber: Faszination des Menschen für das autonomen Schaffen von „Leben“

- 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)
- Die „Schöpfung“ von Frankenstein (aber es entwickelt schlecht für den „Schöpfer“ von Frankenstein, und manchen seiner Familienmitglieder und Freunden werden von der Schöpfung umgebracht)…
- Aber auch in den Bio-Wissenschaften: Cloning, Künstliche Befruchtung, Maschine-Assistierten Reproduktion, etc.
  - Geht das gut?

Golem, “Frankenstein”, HAL 9000



Those above do not look very intelligent, but HAL 9000, who can perform also lip-reading



Gefahren durch Maschinen

- 2001 Space Odyssee
  - HAL 9000 übernimmt Kommando und tötet die mitreisenden Menschen (bis auf einem, der dann HAL „abschaltet“).
  - Interessanter Aspekt: Geschichte der Intelligenz ist auch 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 menschlichen 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“

https://en.wikipedia.org/wiki/ELIZA (13.05.2017)

- **ELIZA** is an early natural language processing computer program created from 1964 to 1966[1] at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum.[2] Created to demonstrate the superficiality of communication between man and machine, Eliza simulated conversation by using a 'pattern matching' and substitution methodology that gave users an illusion of understanding on the part of the program, but had no built in framework for contextualizing events.[3] Directives on how to interact were provided by 'scripts', written originally in MAD-Slip, which allowed ELIZA to process user inputs and engage in discourse following the rules and directions of the script. The most famous script, DOCTOR, simulated a Rogerian psychotherapist and used rules, dictated in the script, to respond with non-directional questions to user inputs. As such, ELIZA was one of the first chatterbots, but was also regarded as one of the first programs capable of passing the Turing Test.

Weizenbaum und Eliza (2)

https://en.wikipedia.org/wiki/ELIZA (13.05.2017)

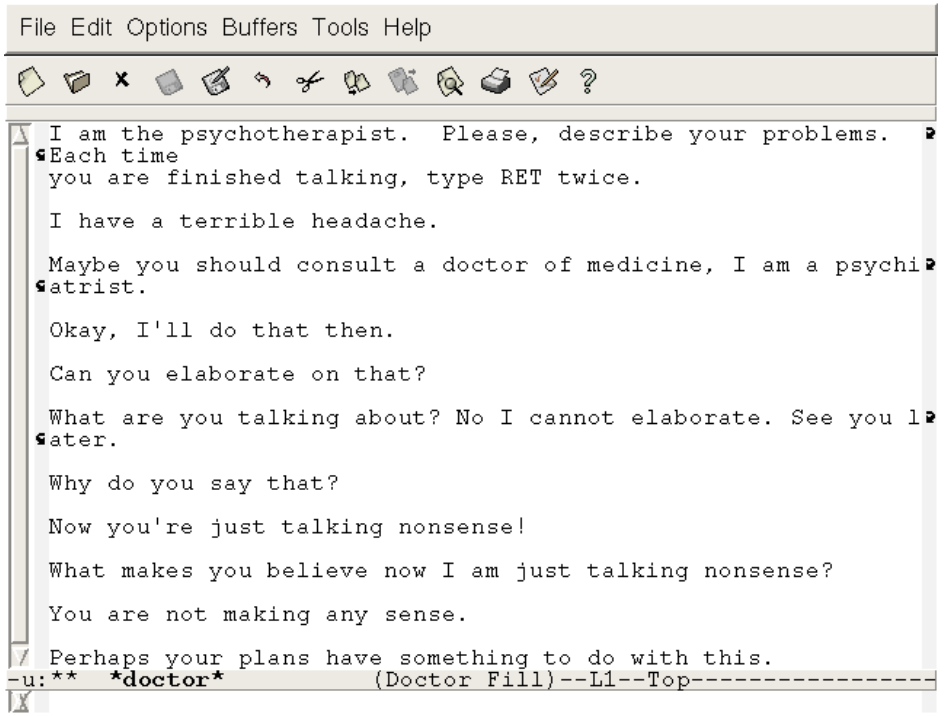
- ELIZA's creator, Weizenbaum regarded the program as a method to show the superficiality of communication between man and machine, but was surprised by the number of individuals who attributed human-like feelings to the computer program, including Weizenbaum's secretary.[2] Many academics believed that the program would be able to positively influence the lives of many people, particularly those suffering from psychological issues and that it could aid doctors working on such patients' treatment.[2][4] While ELIZA was capable of engaging in discourse, ELIZA could not converse with true understanding.[5] However, many early users were convinced of ELIZA's intelligence and understanding, despite Weizenbaum's insistence to the contrary.

Eliza (3)

```
> Hello, I am Eliza.
* I am afraid.
> How long have you been afraid?
* Since midterms.
> Tell me more...
* I am afraid that I will fail
my course.
> Do you enjoy being afraid that
you will fail your course?
```

FIGURE 93. Eliza in action

https://knowledgeillusion.files.wordpress.com/2012/04/4966605.gif



https://upload.wikimedia.org/wikipedia/commons/9/98/GNU_Emacs_ELIZA_example.png

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

- "***Computer Power and Human Reason: From Judgment To Calculation*** (San Francisco: W. H. Freeman, 1976; ISBN 0-7167-0463-3) by Joseph Weizenbaum displays his ambivalence towards computer technology and lays out the case that 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."
  (https://en.wikipedia.org/wiki/Computer_Power_and_Human_Reason, 13.05.2017)

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) Menschen bei Spielen schlagen:
  - Deep Blue ; go Spiel
- 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 ☺