Künstliche Intelligenz

AI Safety

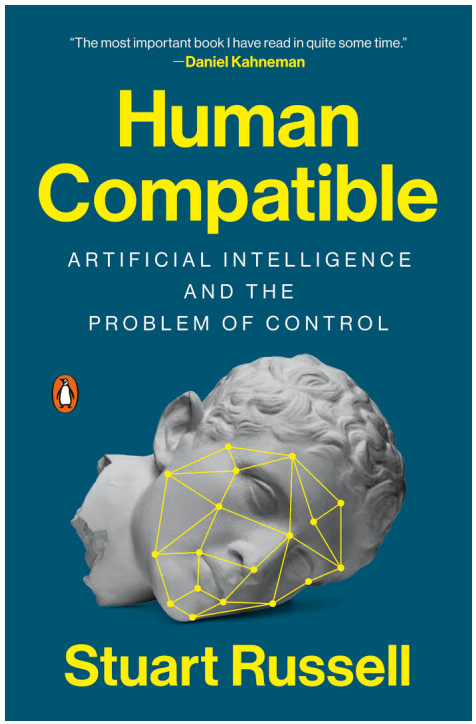
Jun.-Prof. Dr.-Ing. Stefan Lüdtke

Universität Rostock

Institut für Visual & Analytic Computing

Today's Topic

- What we learned about RL so far is insufficient
- How can we develop AL systems which do not kill us eventually?



Progress in AI

Progress in AI

- Rule of thumb: Can automate everything humans can do in < 1 seconds
- Robots acting in the real world ($\rightarrow$ Boston Dynamics)
- Huge progress in visual perception, speec recognition, translation (Deep Learning)
- Even without any more fundamental breakthroughs, current AI technology will have huge impact on:
  - Autonomous Robots on streets and in homes
  - Intelligent assistants (think: Alexa + intelligence)
  - Question Answering, Integration of knowledge and reasoning instead of information retrieval

Progress in AI

- On the other hand: Progress in some fields not as fast as expected
- Current ML methods require tremendous amounts of data
- Training of large models bottlenecked by available data (instead of available compute)
- Not clear whether we can fundamentally solve this problem with current AI technology
- Possible: Failure in some of the promises (autonomous driving) of AI could lead to new AI Winter

Human-Level AI

- What's missing:
- True understanding of language (LLMs are stochastic parrots)
- Integration of Learning and A-Priori-Knowledge ($\rightarrow$ Human brain has lots of priors: Common-Sense-Knowledge, naive physics)
- Planning and Reasoning on large time scales ("How do I need to move my hand to grab a glass" vs. "What do I eat for lunch tomorrow" vs. "Which steps do I need to take for traveling to the Maldives next year")
- Might require conceptual breakthroughs with unclear timeline
- Probably, scale is not all you need

Advantages of Human-Level AI

- Our civilization, wealth almost entirely the result of intelligence
- More intelligence $\rightarrow$ More civilization, wealth, ...

Disadvantages of Human-Level AI

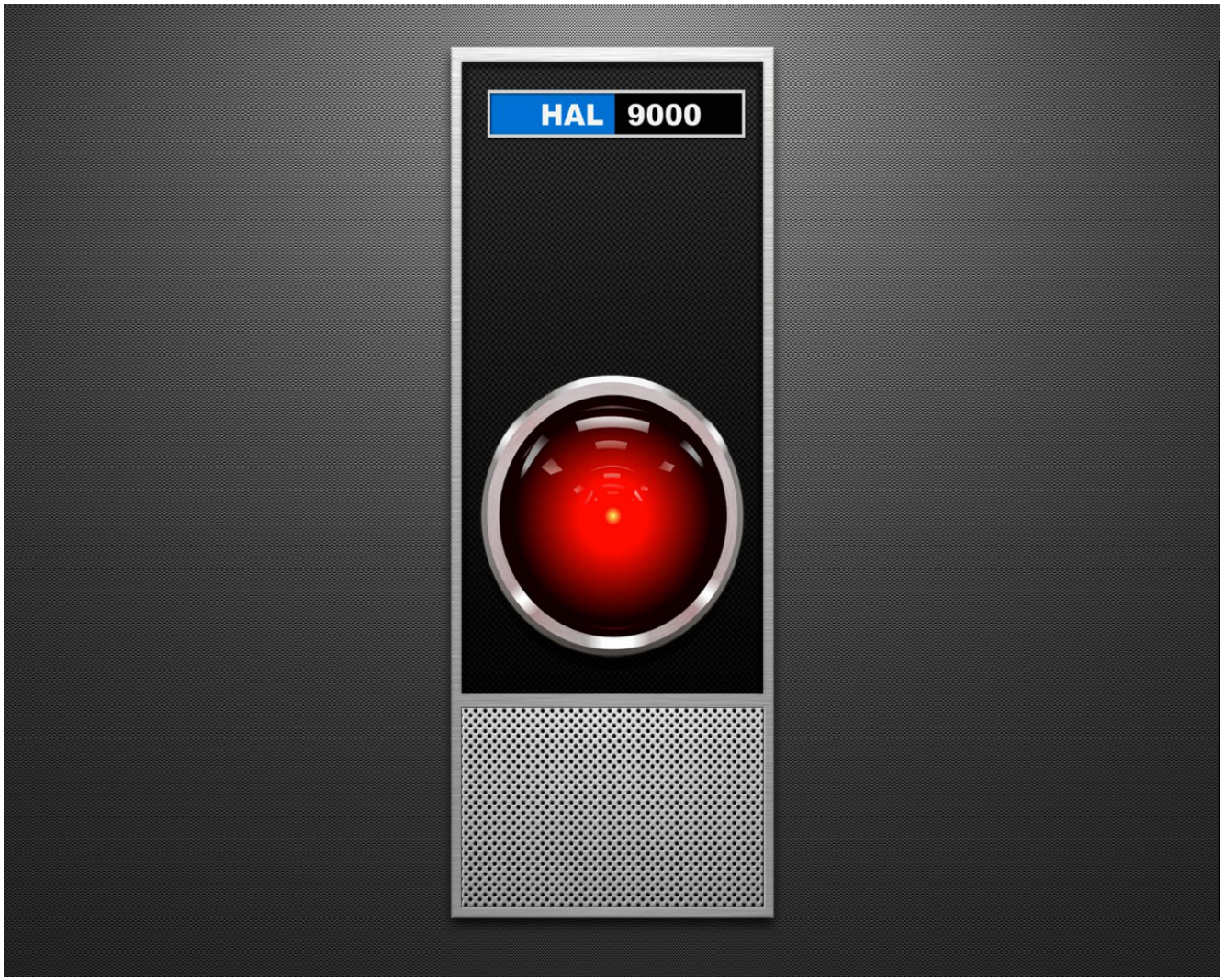
Disadvantages of Human-Level AI

- "*We had better be quite sure that the purpose put into the machine is the purpose which we really desire*" (Norbert Wiener, 1960)
- King Midas problem: Wrong goal
- Off-Switch problem: "You can't fetch the coffee when you're dead"

Paperclip Maximizer

> *Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.*

(Nick Bostrom, 2003)



I'm sorry Dave, I'm afraid I can't do that

What went wrong?

- Human intelligence: Humans are intelligent to the extent that *our* actions lead to achieving *our* goals

What went wrong?

- Human intelligence: Humans are intelligent to the extent that *our* actions lead to achieving *our* goals
- Artificial Intelligence: Machines are intelligent to the extent that *their* actions lead to achieveing *their* goals

What went wrong?

- Human intelligence: Humans are intelligent to the extent that *our* actions lead to achieving *our* goals
- Artificial Intelligence: Machines are intelligent to the extent that *their* actions lead to achieveing *their* goals
  - But: Goals are programmed by us: Costs, reward functions, utility functions, ...

What went wrong?

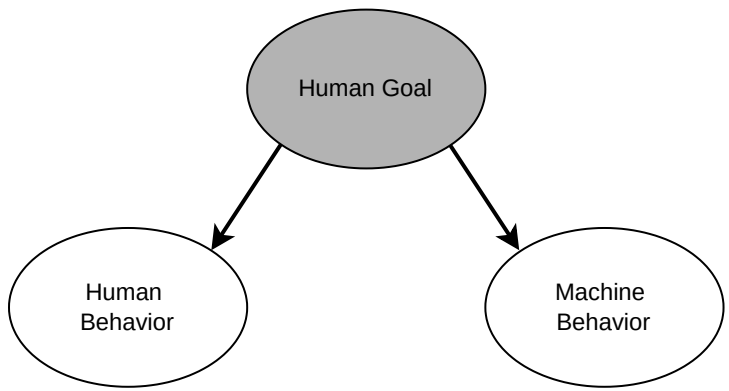
- Human intelligence: Humans are intelligent to the extent that *our* actions lead to achieving *our* goals
- Artificial Intelligence: Machines are intelligent to the extent that *their* actions lead to achieveing *their* goals
  - But: Goals are programmed by us: Costs, reward functions, utility functions, ...
- *Maybe that's a mistake: King Midas problem*
- Instead: Machines are *beneficial* to the extent that *their* actions lead to achieving *our* goals

How does this work: Three ideas

- Only goal of AI is to maximize the realization of human preferences
- The AI is uncertain about these preferences
- The AI gets information about human preferences from human behavior

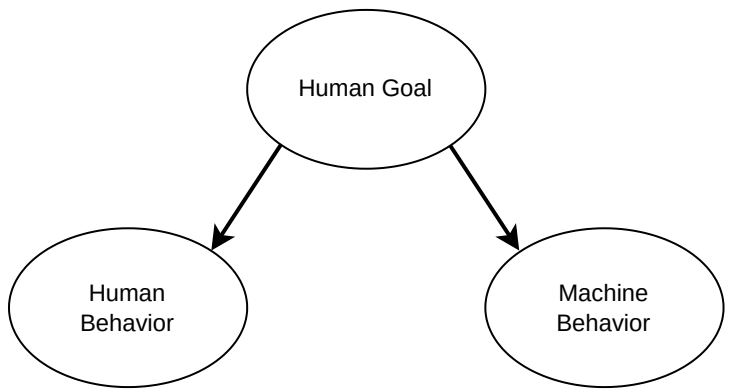
Graphical Model: Current Approach

- Assume that human goal is known (observed)
- Human behavior and machine behavior are independent
- Machine behavior cannot be influenced by human behavior



Graphical Model: New Approach

- Human goal not observed
- Machine needs to infer human goal
- Uses observed human behavior for inference



The Off-Switch Problem, Revisited

- Current AI

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead
  - Therefore, I should disable my off-switch

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead
  - Therefore, I should disable my off-switch
  - (and also kill all customers in Starbucks who are in my way, to fetch the coffee as fast as possible)

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead
  - Therefore, I should disable my off-switch
  - (and also kill all customers in Starbucks who are in my way, to fetch the coffee as fast as possible)
- Beneficial AI

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead
  - Therefore, I should disable my off-switch
  - (and also kill all customers in Starbucks who are in my way, to fetch the coffee as fast as possible)
- Beneficial AI
  - I have to fetch coffee

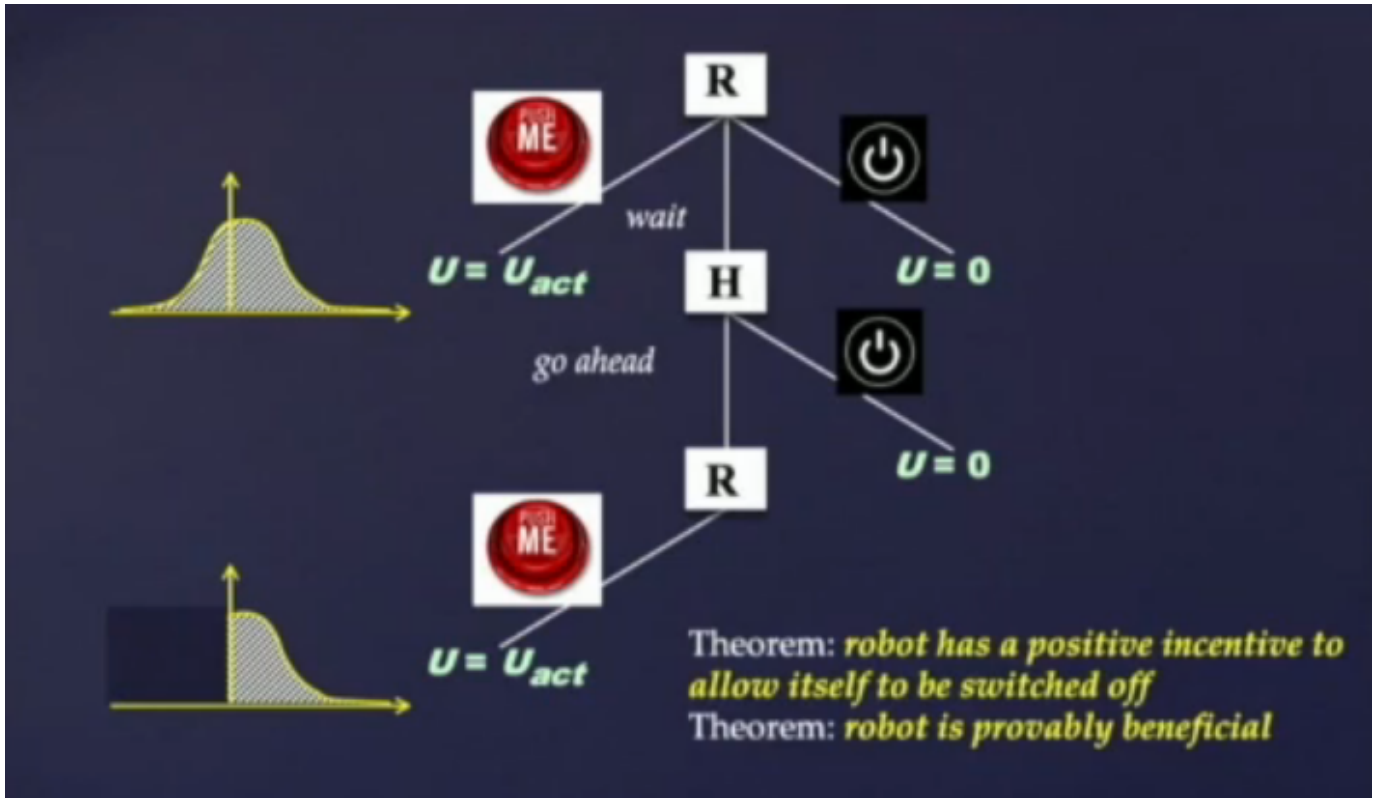
The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead
  - Therefore, I should disable my off-switch
  - (and also kill all customers in Starbucks who are in my way, to fetch the coffee as fast as possible)
- Beneficial AI
  - I have to fetch coffee
  - I might be doing something that is in conflict with human preferences ("negative reward")

The Off-Switch Problem, Revisited

- Current AI
  - I have to fetch coffee
  - I cannot fetch coffee when I'm dead
  - Therefore, I should disable my off-switch
  - (and also kill all customers in Starbucks who are in my way, to fetch the coffee as fast as possible)
- Beneficial AI
  - I have to fetch coffee
  - I might be doing something that is in conflict with human preferences ("negative reward")
  - Therefore, I should allow the human to switch me off, so that the reward can be at least 0.

Assistance Game



Current Research: Humans are no Rational Agents

- Limited computational power
- Emotionally caused behavior
- Uncertainty over own preferences
- Variability of preferences (by machines)

Current Research: Multiple People

- Different preferences
- Individual loyality vs. global utilitarism
- Comparability of inter-personal preferences (same scales?)

Altruismu, Indifference, Sadism

- Assume there are two people, Alice and Bob
- Intrinsic Welfare: w_A and w_B
- Overall happiness: Own welfare + other's welfare:
  - $U_A = w_A + C_{AB}\, w_B$
  - $U_B = w_B + C_{BA}\, w_A$
- Altruismu, Indifference, Sadism depend on C ("caring") factors
- If $C_{AB} = 0$, optimum of $U_A + U_B$ typically gives Alice more intrinsic welfare, but Bob can overall be happier (e.g., Alice is newborn baby of Bob)
- Not clear how to handle $C_{AB} < 0$? Maybe robot should ignore those terms?

Pride and Envy

- Positional goods: Value (= welfare caused by it) of goods depends on comparison to other people's goods
- $U_A = w_A + C_{AB}\, w_B + E_{AB}(w_B - w_A) + P_{AB}(w_A - w_B)$
- $= (1 + E_{AB} + P_{AB})w_A + (C_{AB} - E_{AB} - P_{AB})w_B$
- I.e., pride and envy are mathematically identical to sadism
- But, ignoring these terms might not be sensible, because they seem to be fundamental for human behavior

Summary

- AI is making fast progress and might eventually outperform human capabilities

Summary

- AI is making fast progress and might eventually outperform human capabilities
- Current approach (fixed goals / reward funtions / costs) can lead to unexpected consequences

Summary

- AI is making fast progress and might eventually outperform human capabilities
- Current approach (fixed goals / reward funtions / costs) can lead to unexpected consequences
- Therefore: Allow uncertainty over human preferences: *Provably beneficial AI*

Summary

- AI is making fast progress and might eventually outperform human capabilities
- Current approach (fixed goals / reward funtions / costs) can lead to unexpected consequences
- Therefore: Allow uncertainty over human preferences: *Provably beneficial AI*
- These ideas should become standard for AI developers (bridge engineering always includes safety, and so should AI engineering)