



Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG



CHAI  
Humanities-Centered AI

Ralf Möller, Sylvia Melzer

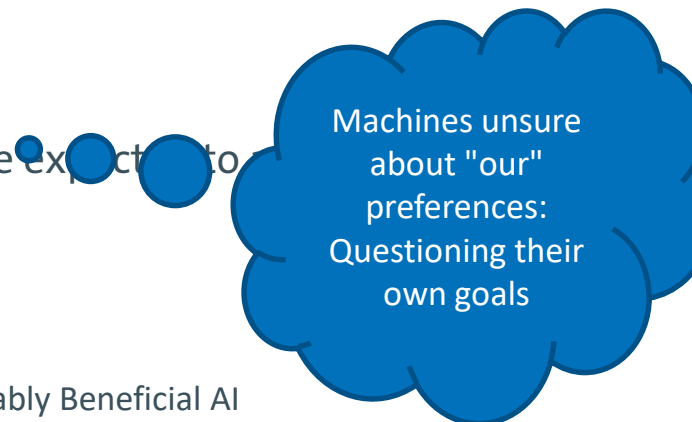
---

# Intelligent Agents - Summary



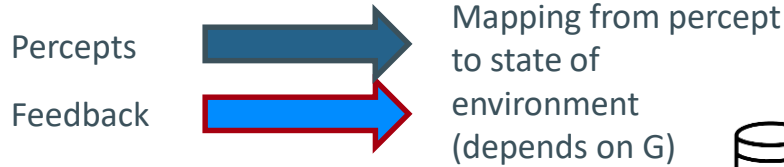
# Where did we go wrong?

- **Humans** are intelligent to the extent that **our** actions can be expected to achieve **our** objectives
- **Machines** are intelligent to the extent that **their** actions can be expected to achieve **their** objectives
  - Give them objectives to optimize (cf control theory, economics, operations research, statistics)
- We don't want machines that are intelligent in this sense
- **Machines** are beneficial to the extent that their actions can be expected to achieve **our** objectives
- We need machines to be provably beneficial

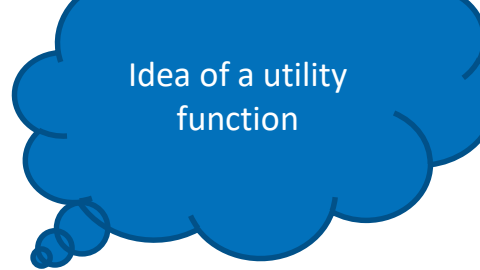
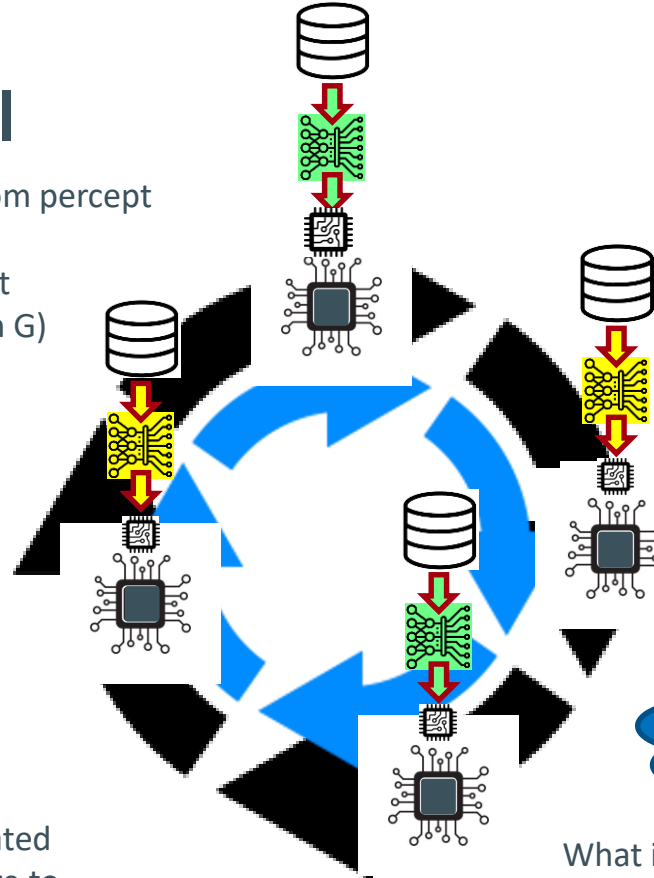


Machines unsure  
about "our"  
preferences:  
Questioning their  
own goals

# Yet some more detail



Return calculated action (prepare to answer the Why question)



Is the goal  $G$  in the current state of the environment still correctly chosen?  
Should I have a new goal  $G'$ ?

My utility will be zero then. How to prevent being switched off?

Hint: Possibly I'll be switched off

Maximize utility (possibly with sequence of actions)

What is the best action in the current state to achieve the goal?  
Strategy for determining action

Representative for dealing  
with unwanted behavior

## Off-Switch Problem

Example: “Fetch some  
coffee”

Agents get better at  
maximizing the built-in  
utility function

What’s bad about better AI?

Can we switch off the agent  
if it “does not work as  
expected”?

“Can’t fetch coffee if I am  
dead.”

Very much underspecified!

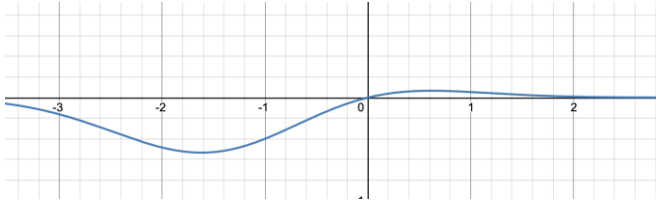
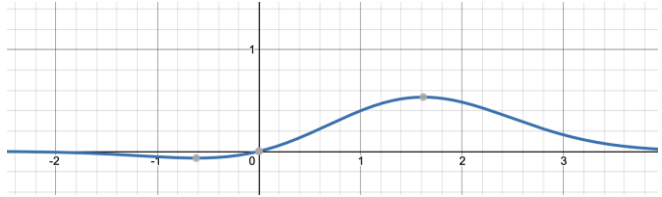
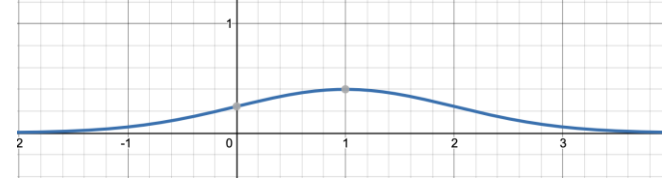
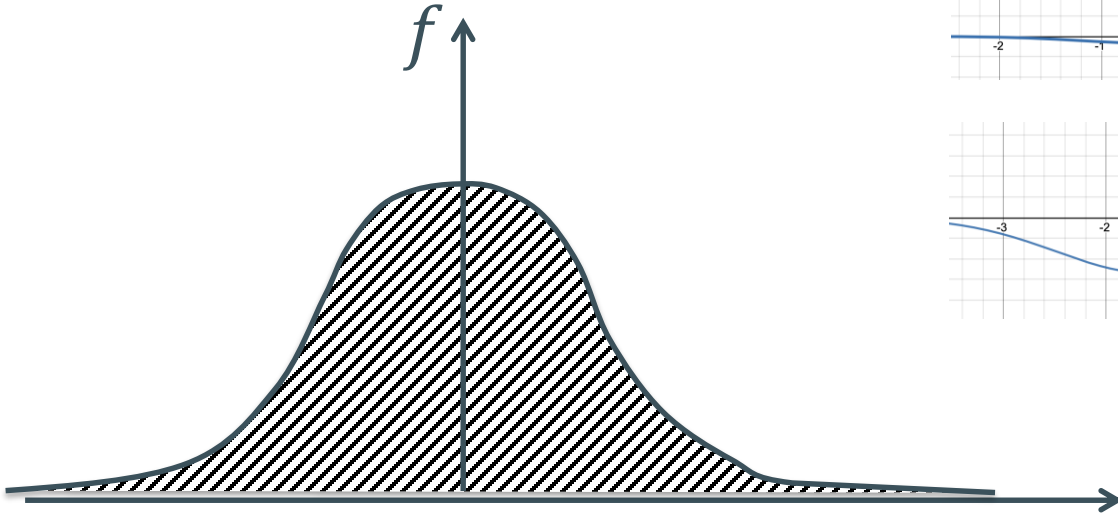
The instruction suggests  
having coffee would have  
higher value than expected a  
priori, ceteris paribus

Incentive for preventing  
others from switching off  
the robot



# Before/after bringing coffee...

- Expected utility  $E[U(state)] = \int x f(x) dx$
- Change the utility function



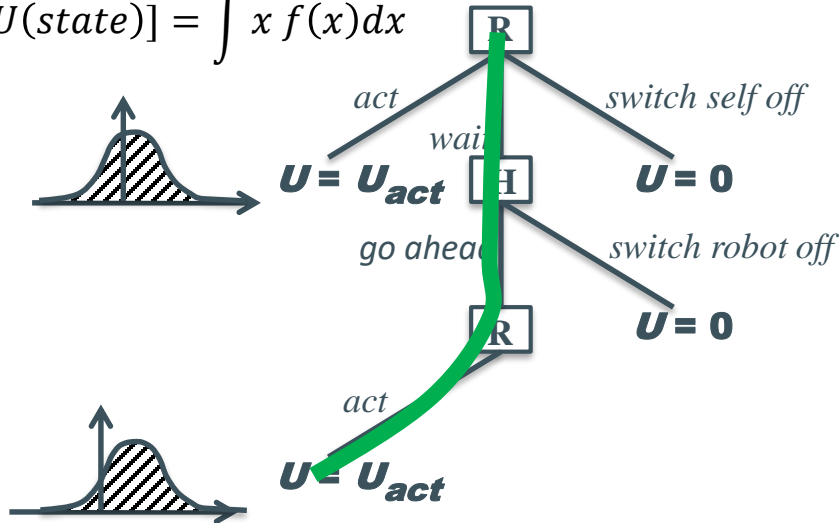
# The off-switch problem

- A robot, given an objective, has an incentive to disable its own off-switch
- Claim: A robot with **appropriate** uncertainty about objective won't behave this way
- Example: Planning for the best action (sequence)

# The off-switch problem

## Example:

$$E[U(\text{state})] = \int x f(x) dx$$



Theorem:

This way robot has a positive incentive to allow itself to be switched off



A white humanoid robot with large, expressive eyes and a friendly appearance. It has a screen on its chest that displays the text 'UHH'. A large red thought bubble is superimposed on the image, containing text about AI. The robot is positioned on the right side of the frame, with its arms slightly outstretched.

Thank you for coming.

AI – Most exciting science today.  
Most relevant for companies.  
Just about to start.

UHH