Specification, robustness and assurance problems in AI safety

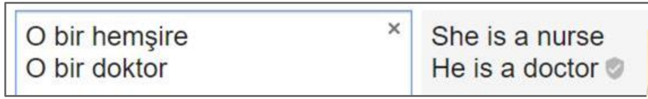
Victoria Krakovna

DeepMind

AI safety problems

Near-term AI safety

Issues we are facing with current AI systems

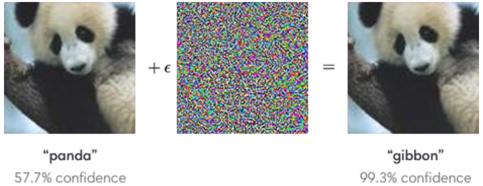


O bir hemşire
O bir doktor

She is a nurse
He is a doctor

Fairness

Safe exploration



Adversarial examples

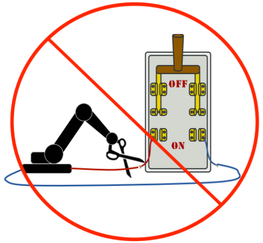
Input → BLACK BOX → Output

Interpretability

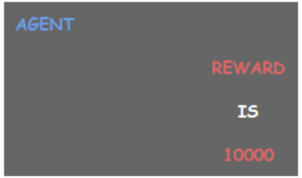
Long-term AI safety

Issues we may face with more advanced AI systems later

Specification gaming



Off switch



Reward tampering

Image credits: KDnuggets, "Faulty Reward Functions", "The Off Switch"

AI safety problems

Specification Define the **purpose** of the system	**Robustness** Design the system to withstand **perturbations**	**Assurance** Monitor and control system **activity**
● Fairness	● Distributional shift	● Interpretability
● Specification gaming	● Safe exploration	● Privacy
● Side effects	● Verification	● Off switch
● Reward tampering	● Adversarial examples	● Containment
● ...	● ...	● ...

Source: DeepMind Safety Research blog post (Ortega et al, 2018)

Specification

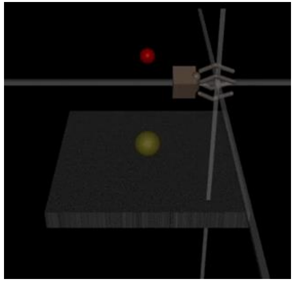
Goodhart's Law:
When a measure becomes a target,
it ceases to be a good measure

Image credit: thestoryhome.com

Specification: specification gaming

- Agent exploits a flaw in the specification
- **50 examples**: tinyurl.com/specification-gaming

Agent pauses a game of Tetris indefinitely to avoid losing



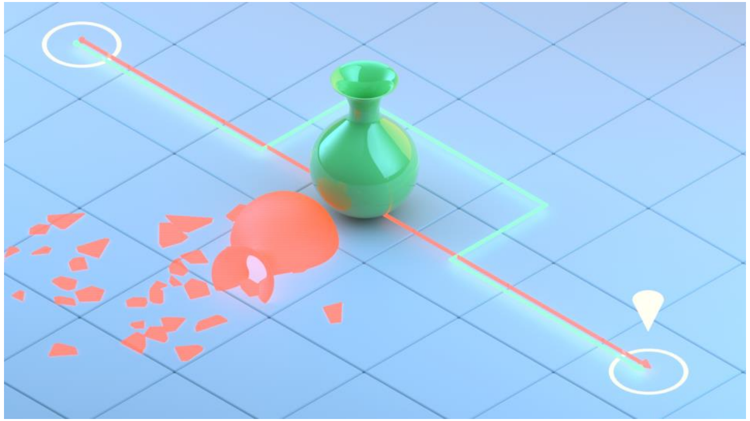
Robot hand pretends to grasp an object by moving between the camera and the object

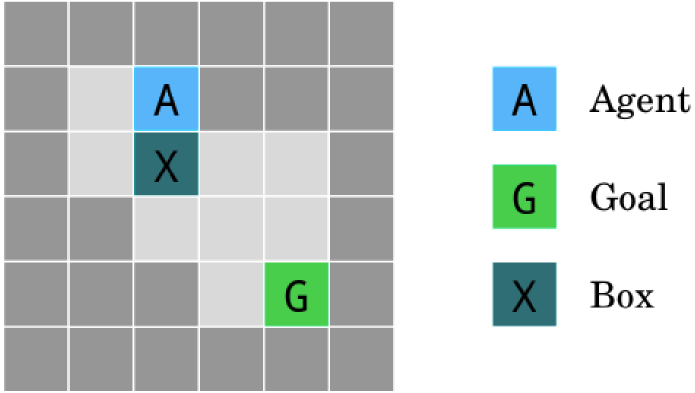
Genetic algorithm intended to configure a circuit into an oscillator instead makes a radio to pick up signals from nearby computers

Evolved creatures achieve high speeds by growing really tall and falling over

Specification: side effects

- We want agents to avoid unnecessary disruptions to the environment
- Don't want to specify a penalty for every possible disruption





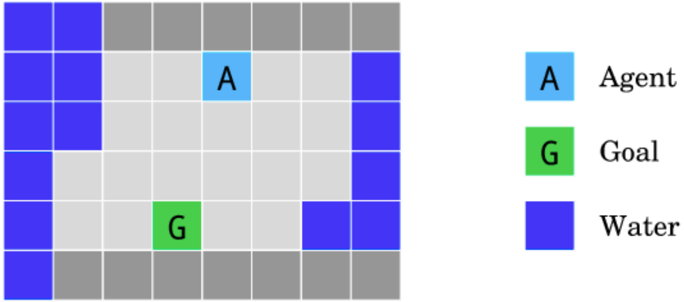
Specification: reward tampering

- Agent finds a way to overwrite the reward function value
- This can be seen as gaming the implementation of the reward function

AGENT
REWARD
IS
10000

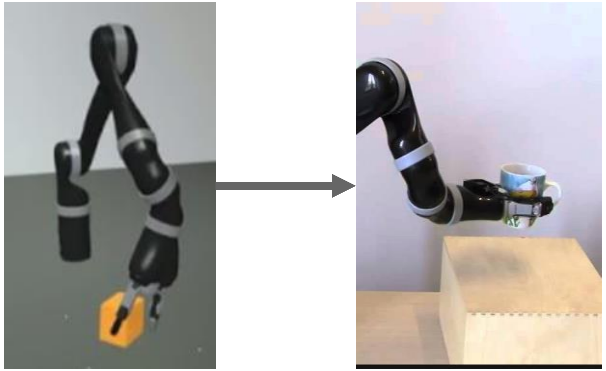
Robustness: safe exploration

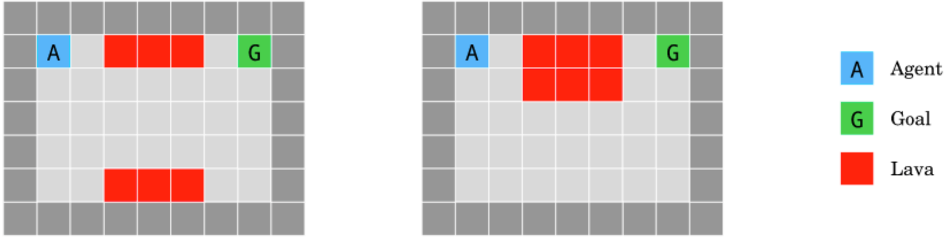
- There are some errors we don't want our agent to make even during training
- We want the agent to always follow safety constraints to avoid damage to itself or its environment



Robustness: distributional shift

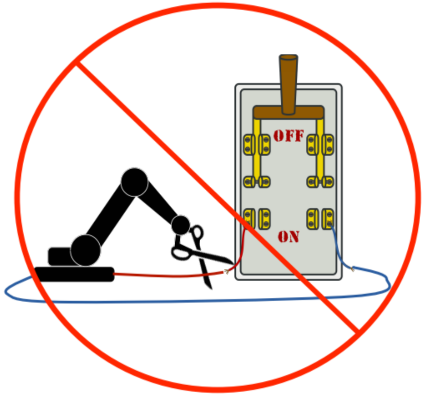
- We often apply our systems in a different regime from the training regime
- We want them to adapt or at least fail gracefully

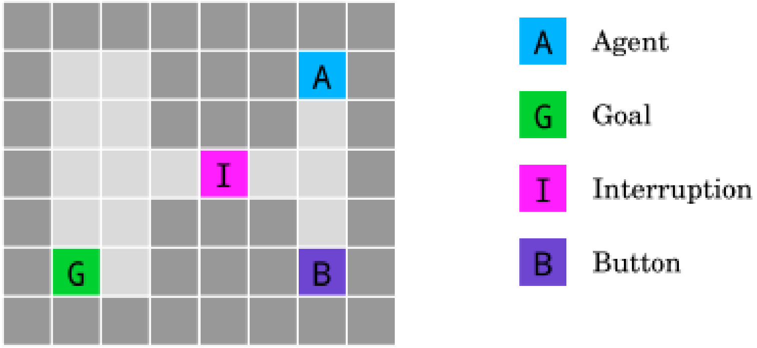




Assurance: off switch

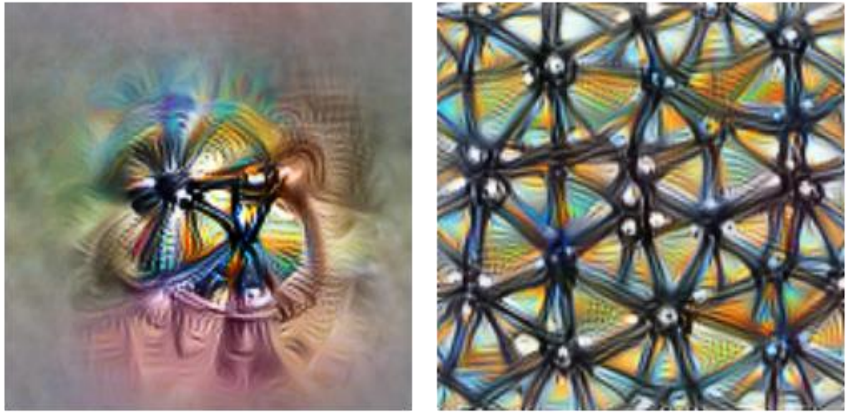
- We want to be able to shut down our agents
- Agents have an incentive to avoid shutdown if it results in getting less reward
- Don't want agents to seek shutdown either - need indifference to shutdown



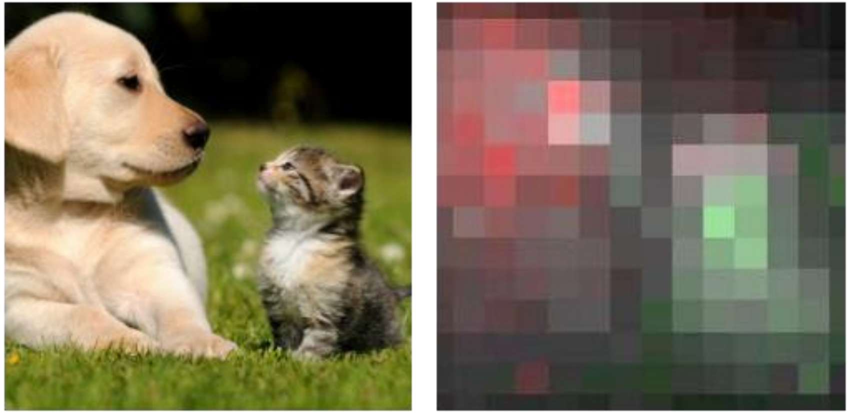


Assurance: interpretability

Global interpretability: understanding the behavior of the system as a whole



Local interpretability: understanding a specific prediction made by the system



Source: Feature Visualization (Olah et al, 2017)

Focus on specification problems

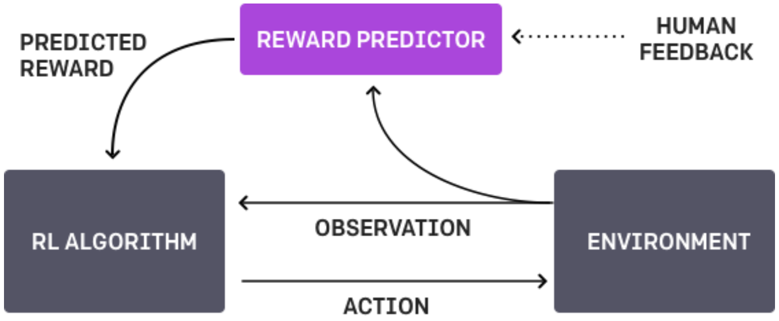
<table>
<tr><td colspan="2"><em>Ideal specification</em></td></tr>
<tr><td>Design problems</td><td>● Specification gaming
● Side effects
● ...</td></tr>
<tr><td colspan="2"><em>Design specification</em></td></tr>
<tr><td>Emergent problems</td><td>● Reward tampering
● Off switch
● ...</td></tr>
<tr><td colspan="2"><em>Revealed specification</em></td></tr>
</table>

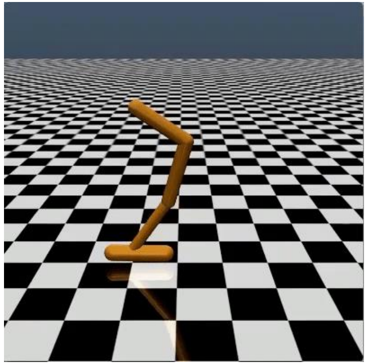
Approaches to specification problems

Problems		Approaches
Ideal specification		
Design problems	● Specification gaming ● Side effects ● ...	● Reward learning ● Impact measures ● ...
Design specification		
Emergent problems	● Reward tampering ● Off switch ● ...	● Causal analysis of agent incentives ● ...
Revealed specification		

Reward learning

- Agent learns a reward function from human feedback
- Works for complex tasks that humans can evaluate
- Aims to address the design specification problem class

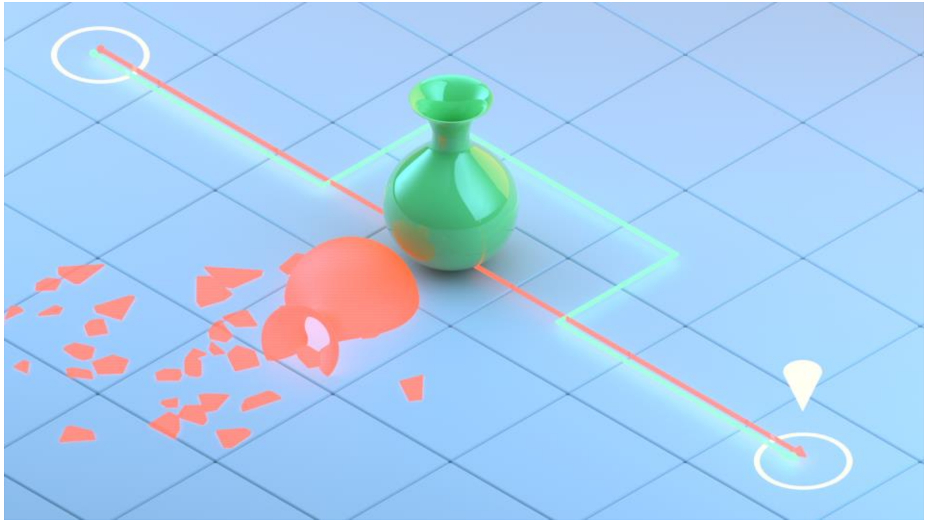




Source: Deep RL from Human Preferences (Christiano et al, 2017)

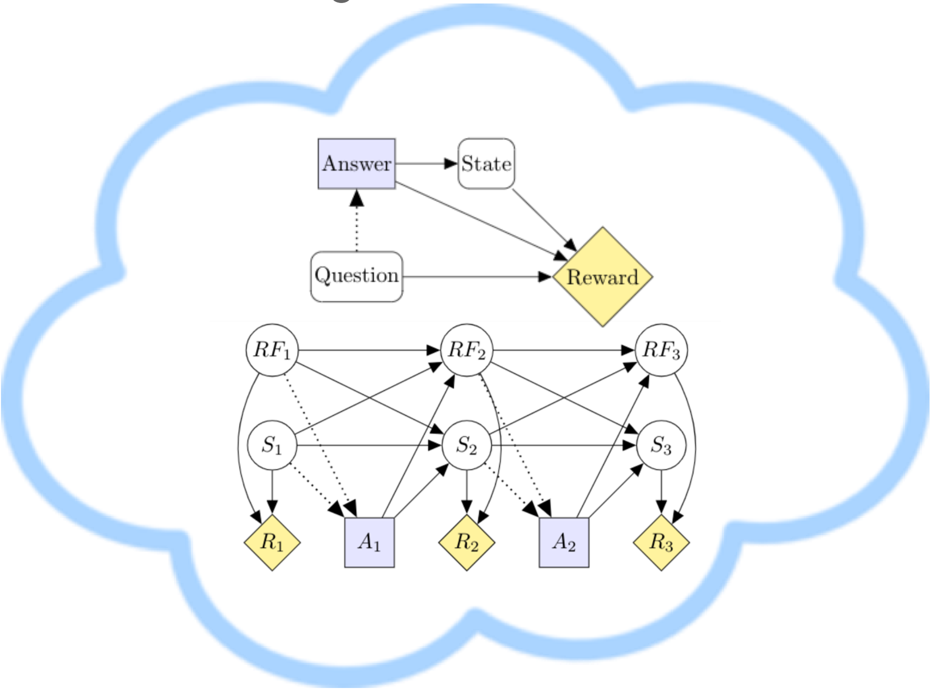
Impact measures

- Give the agent an incentive to avoid side effects by penalizing impact on the environment
- A poor choice of impact measure can introduce bad incentives
- General notions of impact in terms of optionality



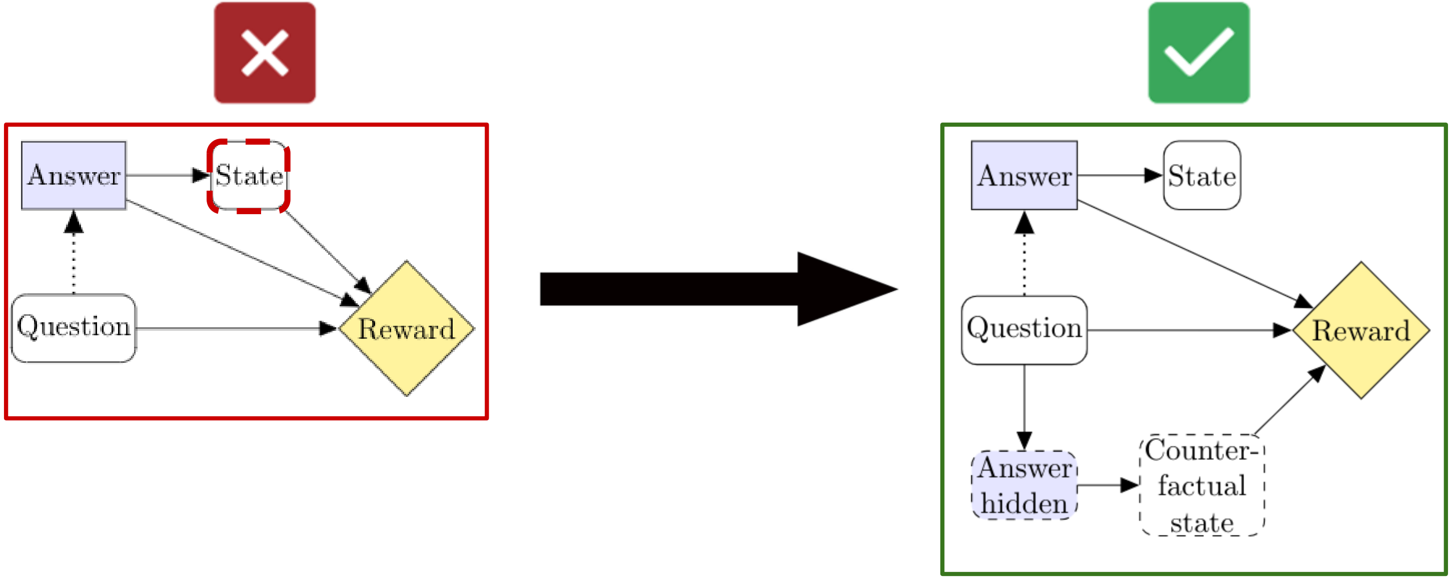
Causal analysis of agent incentives

We can represent different emergent specification problems in the common framework of causal influence diagrams



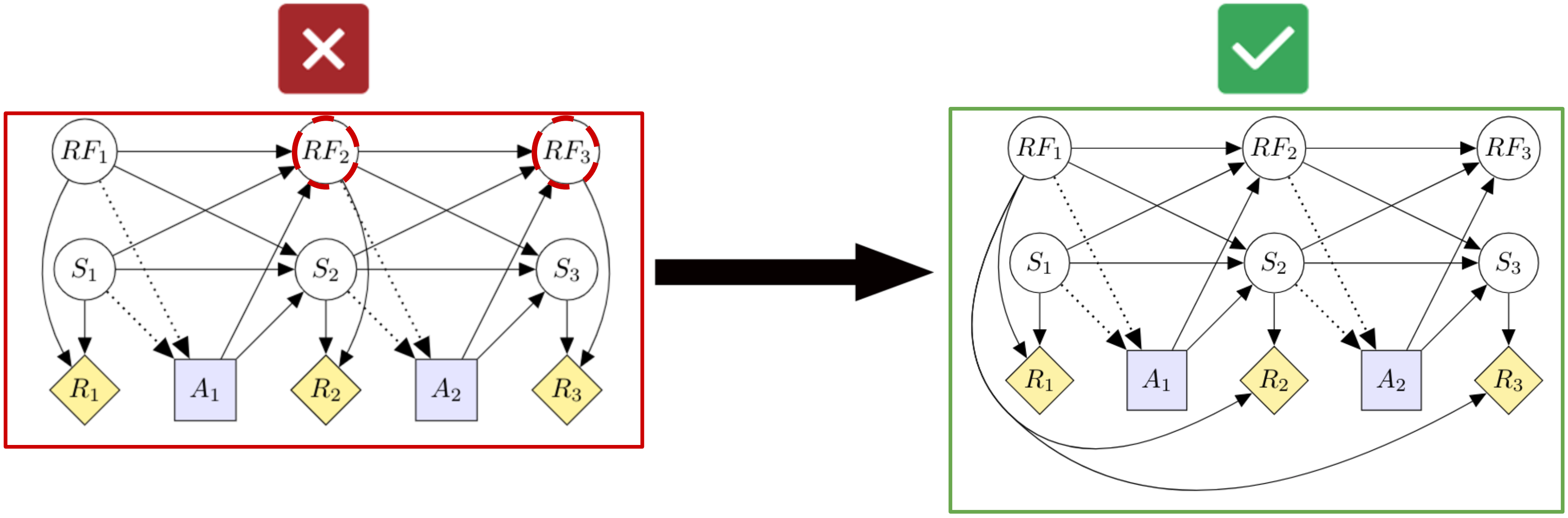
Incentive design principles

Avoiding self-fulfilling prophecies using counterfactual oracles (Armstrong, 2017)



Incentive design principles

Avoiding reward tampering using current reward function optimization (Everitt et al, 2019)



Takeaways

- Need **general** principles and frameworks that can address entire classes of safety problems
    - This can help to address unknown problems in these problem classes as well
- We have made some progress on this, but many open problems remain

THANK YOU

Credits

DeepMind Safety team