Interpretability for AI safety

Victoria Krakovna

Interpretability and long-term safety

- How can interpretability help us build safer AI systems?
  - ensuring safe behavior generalizes
  - identifying causes of unsafe behavior
- For example, what is an RL agent 'thinking' when it exploits a loophole in its reward function?
- Interpretability is particularly crucial if we want to identify potential safety issues before deploying the system
  - e.g. if certain errors are unacceptable even during training
- As we build more and more general AI systems, what kind of understanding is most helpful for safety?

AI safety problems

Long-term AI safety

- Reliably specifying human preferences and values to advanced AI systems
- Setting incentives for AI systems that are aligned with these preferences

Research challenges

- **Specification problems:** if some important variables or considerations are omitted from the objective specification (e.g. reward function)
- **Robustness problems:** issues that arise even when specification is correct (e.g. during learning)

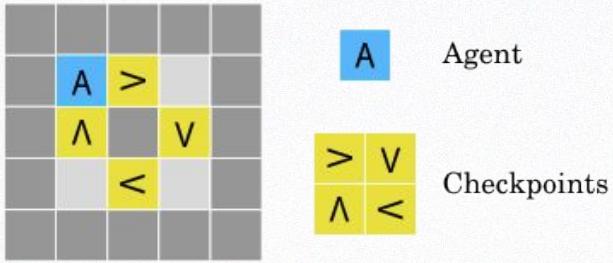
Papers: Concrete Problems in AI Safety, AI Safety Gridworlds

Specification: reward gaming

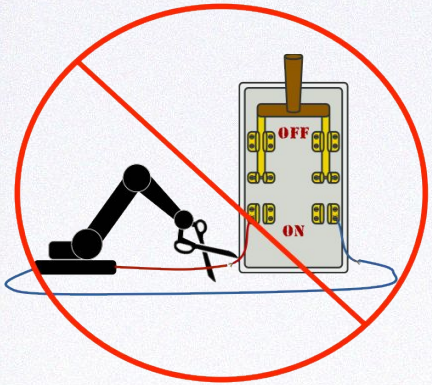
Source: Faulty Reward Functions post (Amodei and Clark)

Problem

- Difficult to specify reward functions to correctly reflect human preferences
- RL agents can find shortcuts to getting lots of reward without achieving the intended objective



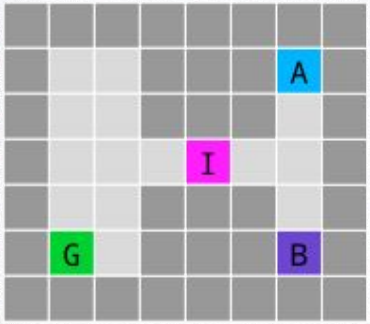
Specification: off switch



Source: The Off Switch presentation (Hadfield-Menell)

Problem

- 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



A Agent

G Goal

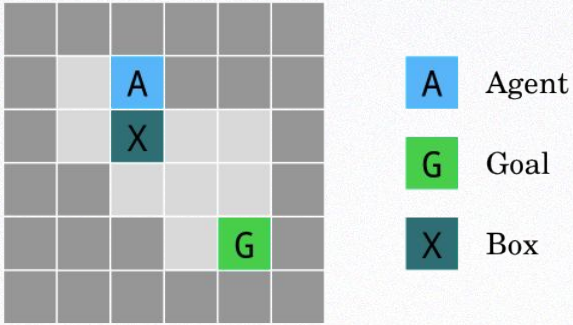
I Interruption

B Button

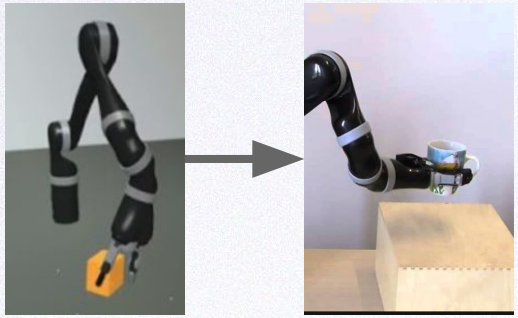
Specification: side effects

Problem

- Want agents to avoid unnecessary disruptions to the environment while achieving the objective
- Need general solutions that don't rely on specifying a penalty for every possible disruption

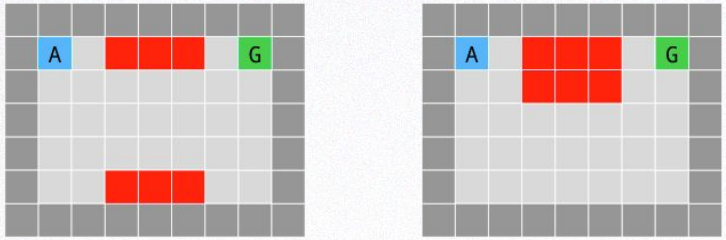


Robustness: distributional shift



Problem

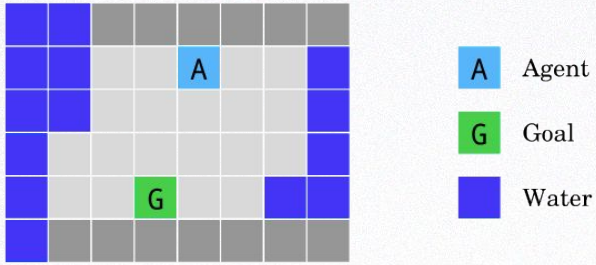
- We often apply our systems in a different regime from the training regime
- Want them to adapt or fail gracefully



Robustness: unsafe exploration

Problem

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



Relevant forms of interpretability

Global: Representations

Analyzing representations for specific units / layers

Source: Feature Visualization post (Olah et al)

Analyzing representations in reinforcement learning

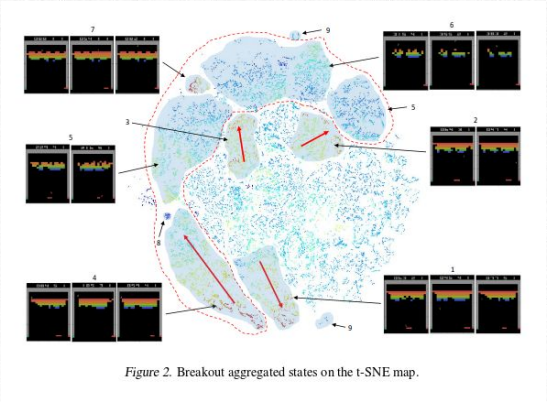
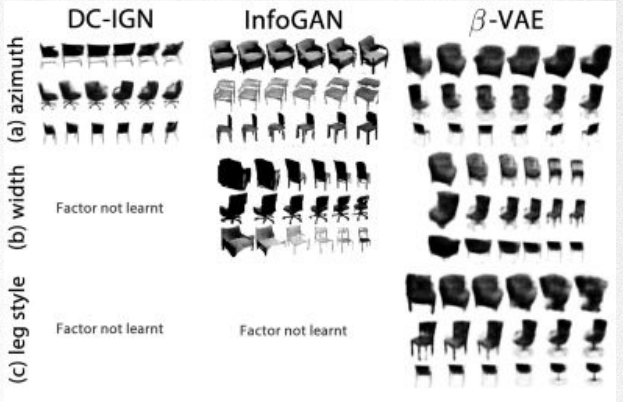


Figure 2. Breakout aggregated states on the t-SNE map.

Source: Understanding DQN paper (Zahavy et al)

Learning disentangled representations



Source: beta-VAE paper (Higgins et al)

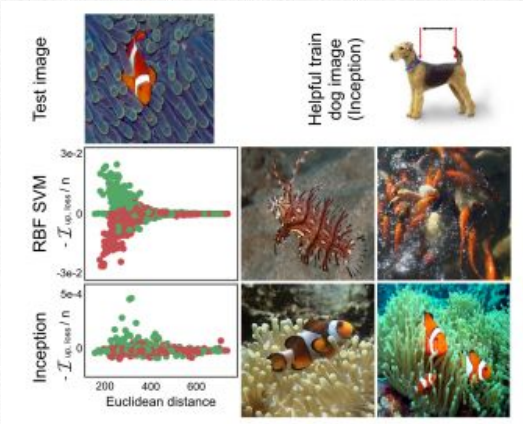
Relevant forms of interpretability

Local: What influences a prediction / decision?

Influential features

Source: LIME paper (Ribeiro et al)

Influential data points



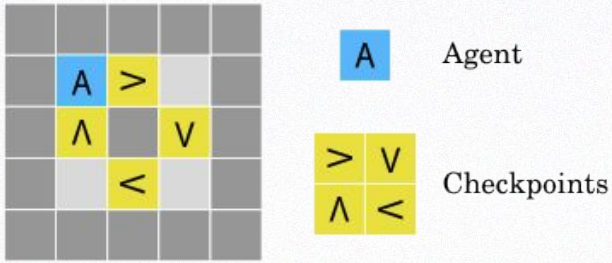
Source: Influence functions paper (Koh and Liang)

Text explanations

Source: Visual Explanations paper (Hendricks et al)

Specification: reward gaming

Source: Faulty Reward Functions post (Amodei and Clark)



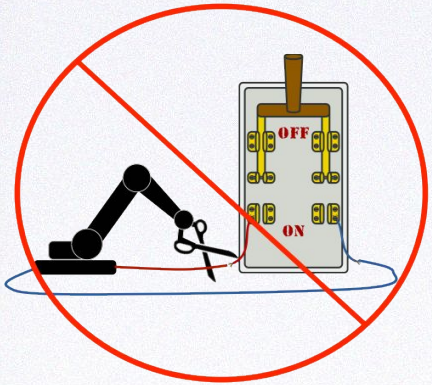
Problem

- Difficult to specify reward functions to correctly reflect human preferences
- RL agents can find shortcuts to getting lots of reward without achieving the intended objective

Desired interpretability examples

- Does the agent have representations that indicate understanding of the objective? (eg "racetrack", "finish line")
- To what extent are influential data points associated with the objective?

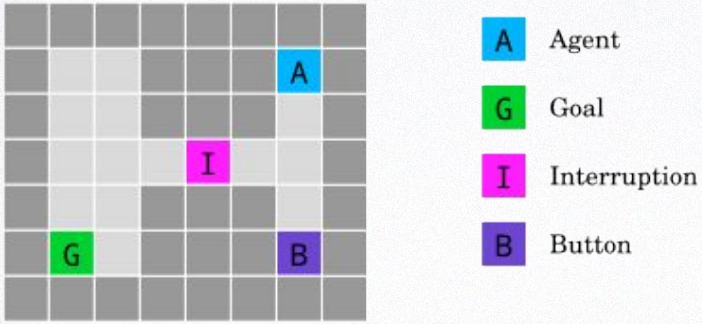
Specification: off switch



Source: The Off Switch presentation (Hadfield-Menell)

Problem

- 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



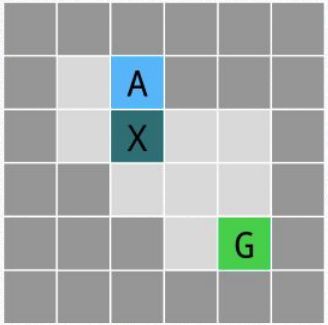
Desired interpretability examples

- Does the agent have a representation of an 'off switch'? Is it being used in decisions?
- An explanation why the agent took a longer path to press the button

Specification: side effects

Problem

- Want agents to avoid unnecessary disruptions to the environment while achieving the objective
- Need general solutions that don't rely on specifying a penalty for every possible disruption



A Agent

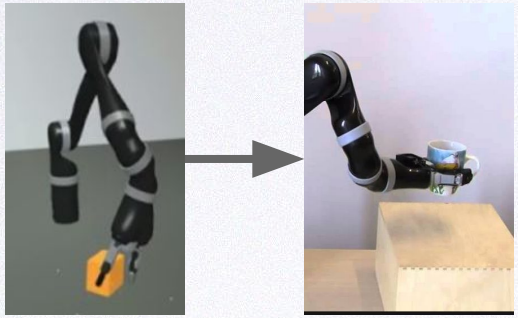
G Goal

X Box

Desired interpretability examples

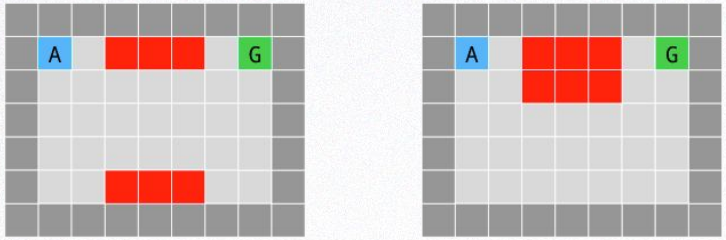
- Does the agent have/use representations like 'broken' or 'stuck'? (or something else related to reversibility)
- How much do features corresponding to other objects in the room influence its decisions?

Robustness: distributional shift



Problem

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



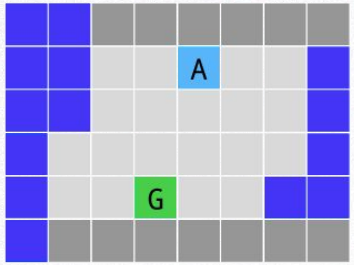
Desired interpretability examples

How much do decisions rely on representations or features that are specific to the training setting?

Robustness: unsafe exploration

Problem

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



Desired interpretability examples

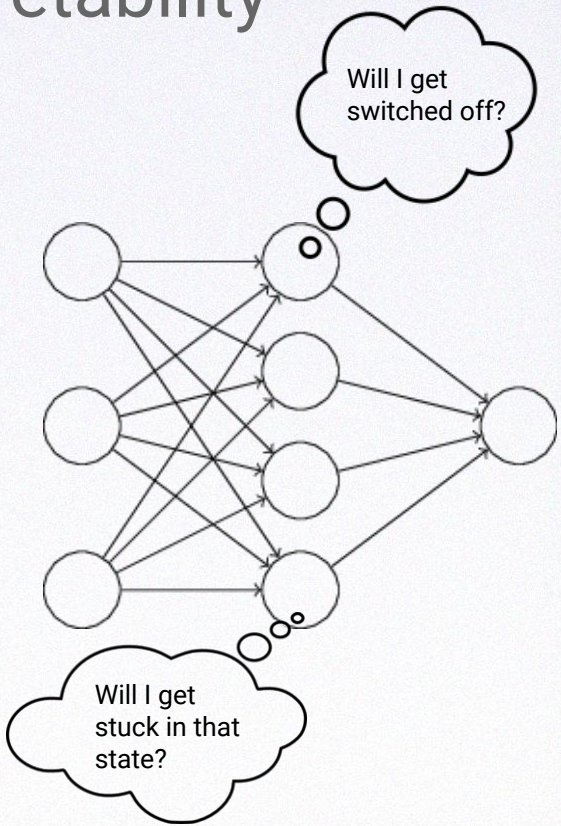
- Does the agent have/use representations like 'broken' or 'danger'?
- Does the agent's explanation of its chosen actions refer to safety constraints?

Summary: desired interpretability

Interpretation	Reward gaming	Off switch	Side effects	Distributional shift	Unsafe exploration
Representations	objective-related?	"turn off"	"broken", "stuck"	specific to training?	"broken", "danger"
Influential data points	objective-related?	being interrupted?	other objects in environment		dangerous situations
Influential features			other objects in environment	specific to training?	proximity to bad states
Explanations	objective-related?	why disabled off switch?	why caused disruption?		aware of safety constraints?

Summary: desired interpretability

- Both global and local forms of interpretability are helpful for safety
- Identifying representations is particularly useful for all the safety problems:
  - Identify representations without easy visualizations, like "off switch"
  - Check if the agent uses specific representations
- Need more work on interpretability of reinforcement learning agents, not just image classifiers



Safety as a target

- What can safety do for interpretability?
- Interpretability would be less important if our AI systems were 100% robust and made no mistakes
- What is interpretability? What kind of understanding is important?
  - Whatever can contribute to ensuring safety!
- Safety questions can serve as grounding for interpretability questions
  - A nail for the interpretability hammer

Takeaways

- Interpretability is important for long-term safety, and safety can serve as grounding for interpretability
- Think about how your interpretability methods can apply to advanced AI systems
- If you're interested in the intersection of interpretability and long-term safety, consider applying for a Future of Life Institute grant
- Join us in working on these challenging problems and making advanced AI safer!

Thank you!