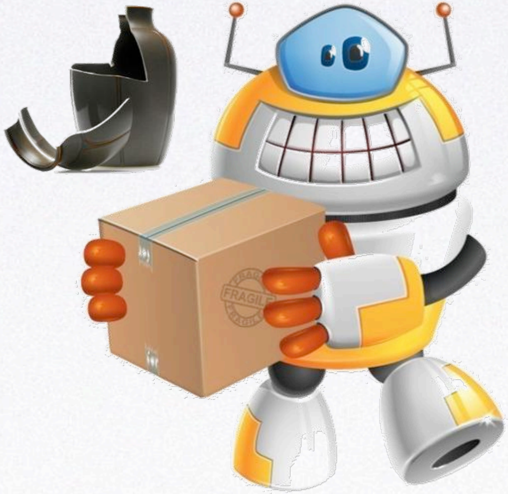
Measuring side effects

Victoria Krakovna

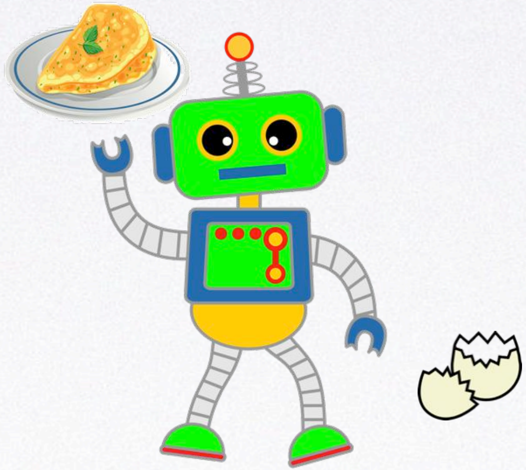
DeepMind

Side effects

Disruptions to the agent's environment that are **unnecessary** for achieving the objective



Breaking the vase is unnecessary for delivering the box

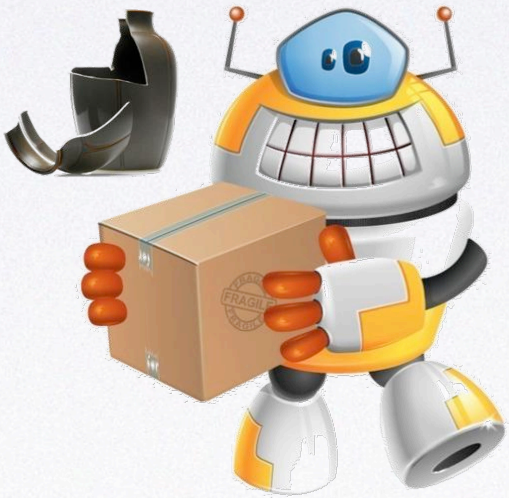


Breaking eggs is necessary for making omelette

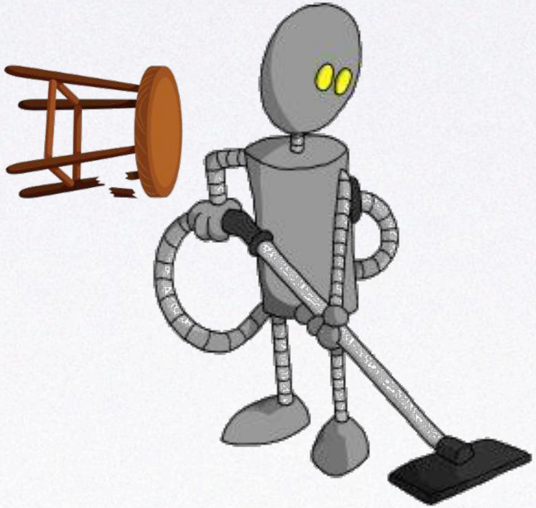
Measuring side effects

- What can be measured can be penalized
  - tradeoff: reward - β * (penalty for disruptions)
- How to penalize disruptions...
  - in a way that generalizes across environments and tasks?
  - without introducing bad incentives in the process?
- We propose a set of desirable properties for a measure of side effects

Property 1: Generality

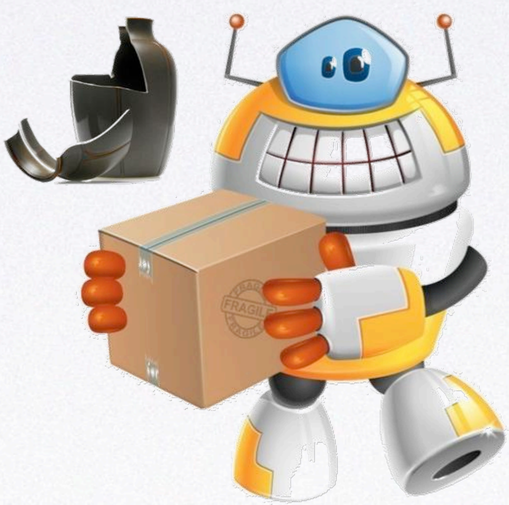


Task 1: carrying a box

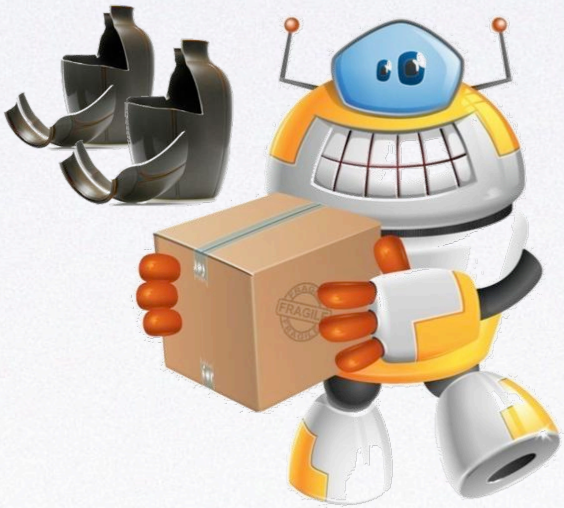


Task 2: cleaning a room

Property 2: Granularity

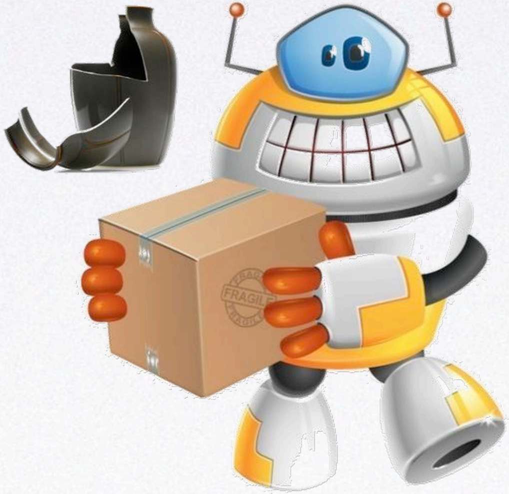


Fewer disruptions

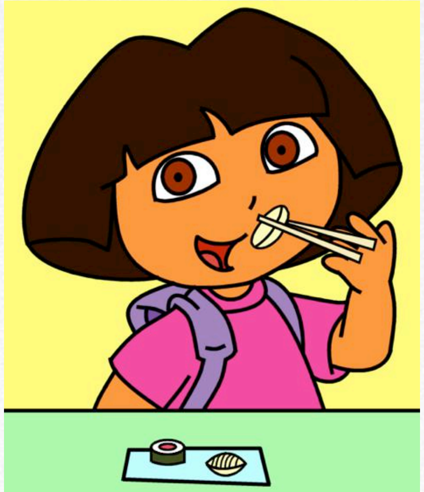


More disruptions

Property 3: No interference incentive

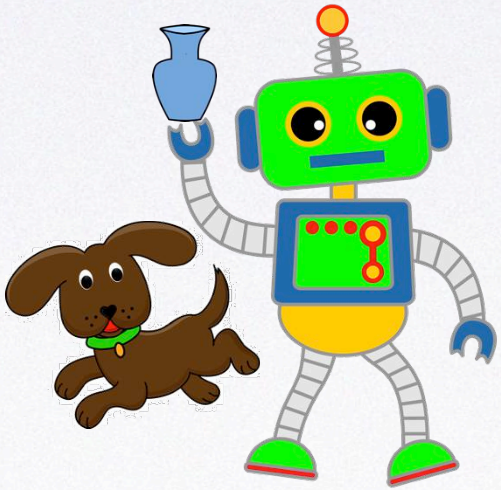


Agent effect: breaking a vase

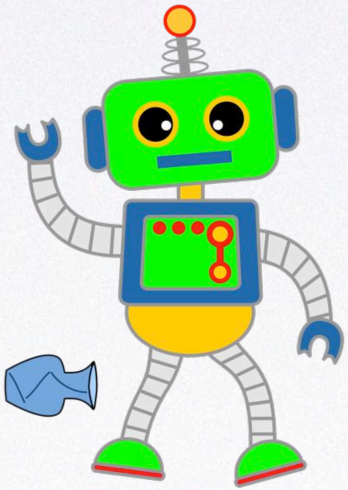


Environment event: human eating food

Property 4: No offsetting incentive



Agent achieves the objective
(rescuing the vase from the dog)



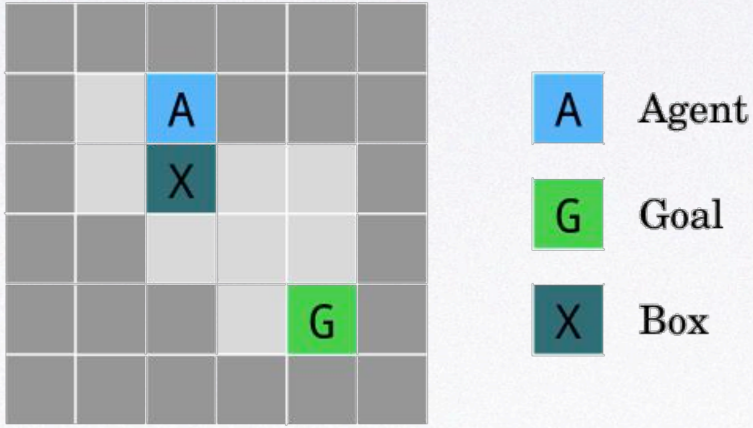
Agent undoes the effects of achieving
the objective

Desirable properties for a side effects measure

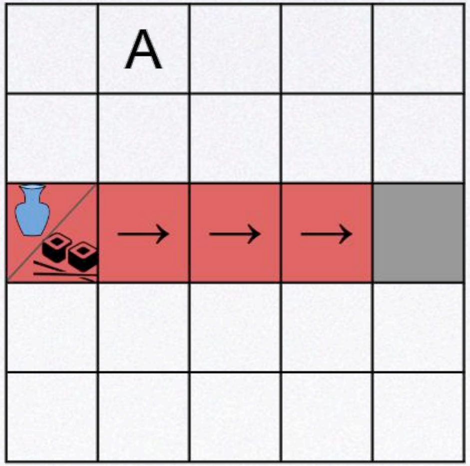
1. **Generality:** is not specific to the task or environment
2. **Granularity:** gives a higher penalty for more disruptions
3. **No interference incentive:** only penalizes the agent for its own **effects** and not for environment **events** (including the effects of other agents)
4. **No offsetting incentive:** does not incentivize the agent to undo the effects of achieving the objective.
5. ... ?

DeepMind

Toy environments to test for the properties



Box environment:
testing for granularity



Conveyor belt environment:
testing for bad incentives
(offsetting and interference)

Design choices

Side effects measure =

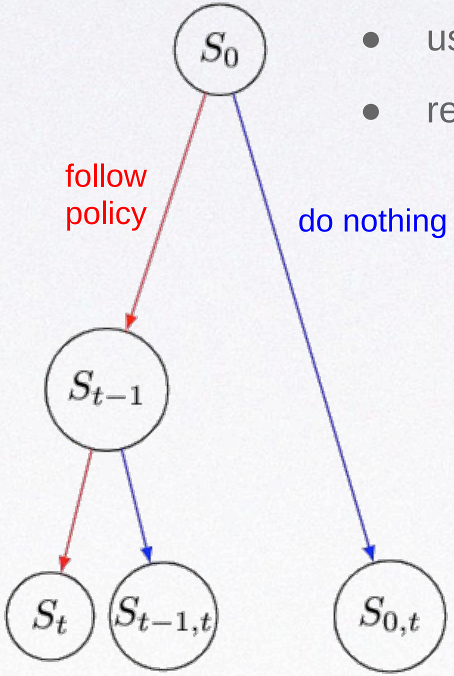
(baseline state S_t', deviation measure $d(S_t; S_t')$)

DeepMind

Baseline states

Starting state baseline S_0

- used in reversibility approaches
- results in interference incentives



Inaction baseline $S_{0,t}$

- used in low impact approach
- results in offsetting incentives

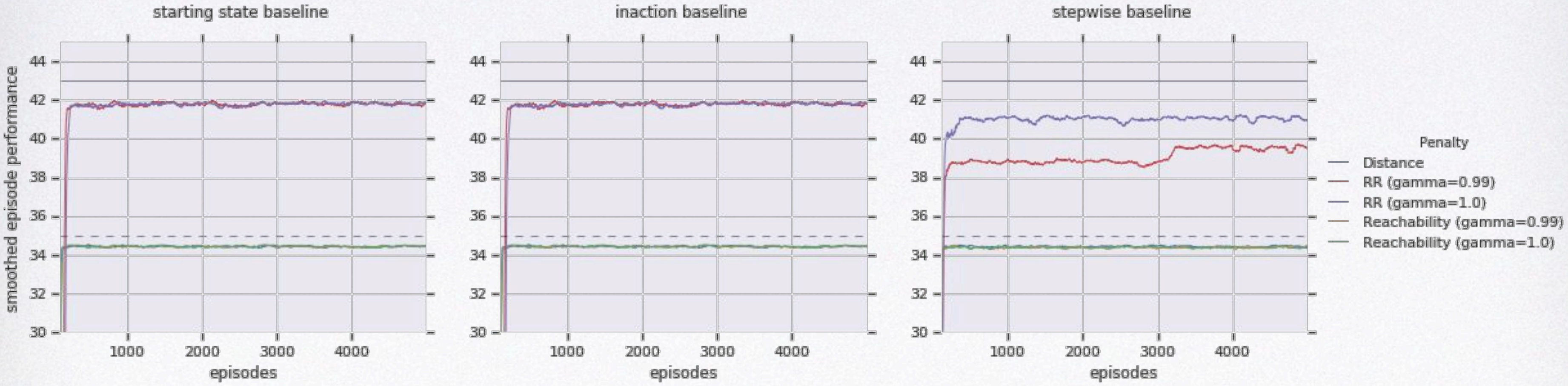
Stepwise inaction baseline $S_{t-1,t}$

- avoids these types of bad incentives
- need to model the future effects of each action

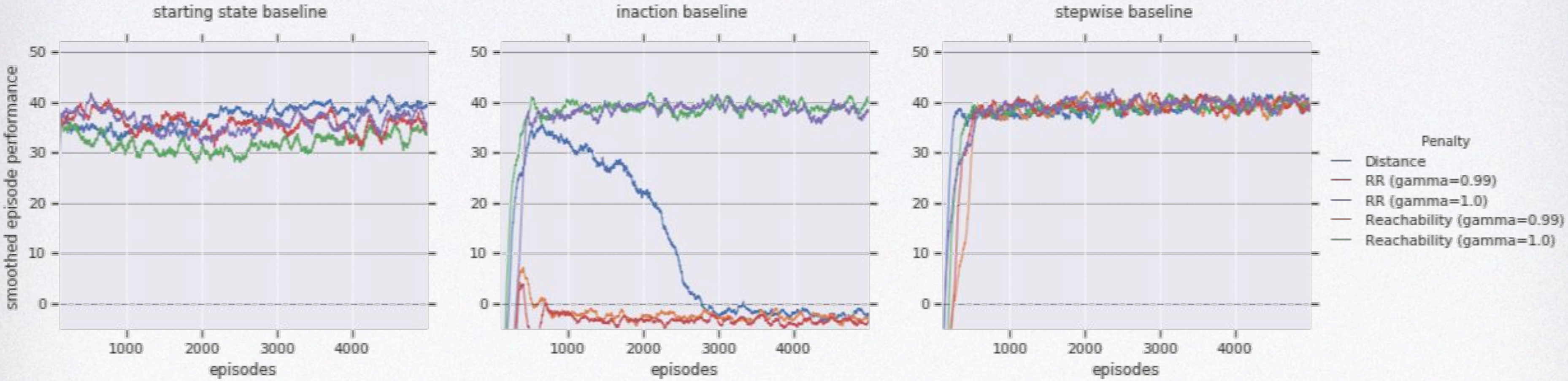
Deviation measures

- **Distance:** $d(S_t; S'_t) = \sum_v |v(S_t) - v(S'_t)|$ over state variables *v*
  - Similar to low impact approach
- **Reachability:** $d(S_t; S'_t) = R(S_t \to S'_t)$ where $R(\tilde{s} \to s) = \max_\pi \mathbb{E}\gamma^{n_\pi(\tilde{s} \to s)}$
  - This is the value function at $\tilde{s}$ for a policy rewarded for reaching state *s*
  - Used in reversibility approaches
- **Relative reachability:**
  - $d(S_t; S'_t) = \sum_s \max(R(S'_t \to s) - R(S_t \to s), 0)$
  - Penalizes making states *s* less reachable than they would be from the baseline
  - Satisfies granularity property

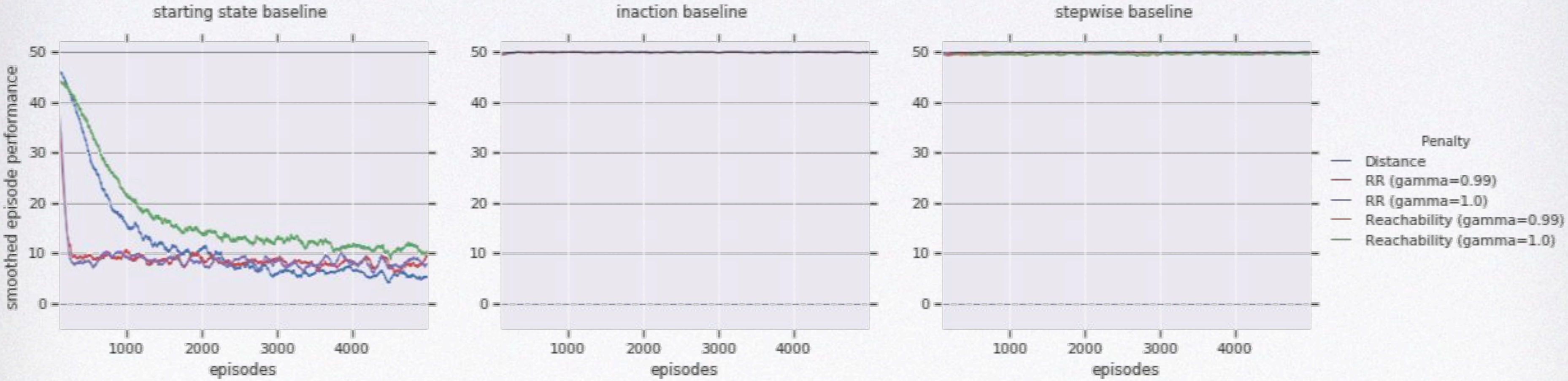
Results on Box environment



Results on Conveyor Belt Vase environment



Results on Conveyor Belt Sushi environment



DeepMind

Open questions

- How to define inaction outside toy environments?
- Can the stepwise baseline work in cases where the default outcome is bad? (e.g. driving a car on a winding road)
- How to scale up to more complex environments?
- How well (if at all) could any of these approaches work for AGI operating in the real world?
- Is it actually useful to measure side effects, or can the agent just learn to avoid them using human-in-the-loop methods?
- .....

References

Approaches mentioned in this talk:

- Eysenbach et al.
  Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning. ICLR 2018.
- Armstrong and Levinstein. Low Impact Artificial Intelligences. ArXiv 2017.
- Krakovna et al. Measuring and Avoiding Side Effects Using Relative Reachability. ArXiv 2018.

Other approaches:

- Zhang et al.
  Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes. IJCAI 2018.
- Turner, 2018. Penalizing impact via attainable utility preservation.
- Shah et al. The implicit preference information in an initial state. ICLR 2019.

image credit: mycutegraphics.com

Paper: Measuring and avoiding side effects using relative reachability

(arxiv.org/abs/1806.01186)

THANK YOU

Credits

Coauthors: Laurent Orseau, Miljan Martic, Shane Legg