Block 3: AI Safety Applications

Tom Everitt

July 10, 2018

Table of Contents

Motivation

What if we succeed?

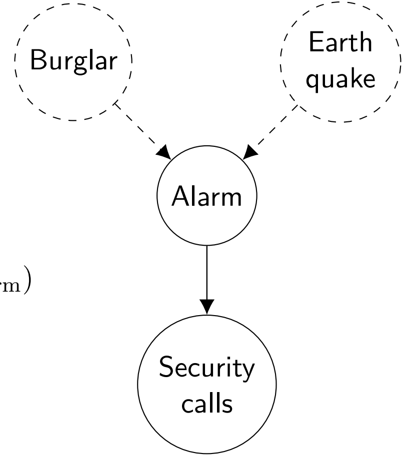
Motivation

What if we succeed?

Extensions of the UAI framwork enable us to:

- Formally model many safety issues
- Evaluate (combinations of) proposed solutions

Causal Graphs



Structural equations model:

$$
\begin{aligned}
\text{Burglar} &= f_{\text{Burglar}}(\omega_{\text{Burglar}}) \\
\text{Earthquake} &= f_{\text{Earthquake}}(\omega_{\text{Earthquake}}) \\
\text{Alarm} &= f_{\text{Alarm}}(\text{Burglar}, \text{Earthquake}, \omega_{\text{Alarm}}) \\
\text{Call} &= f_{\text{Call}}(\text{Alarm}, \omega_{\text{Call}})
\end{aligned}
$$

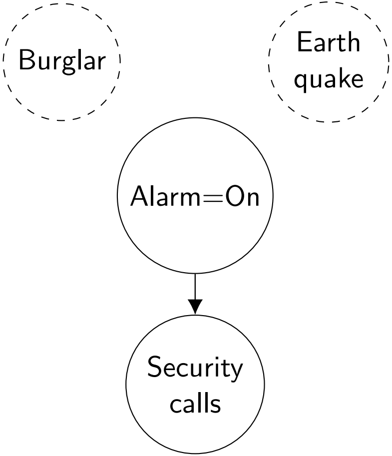
Factored probability distribution:

$$
\begin{aligned}
&P(\text{Burglar}, \text{Earthquake}, \text{Alarm}, \text{Call}) \\
&\qquad = P(\text{Burglar})P(\text{Earthquake})P(\text{Alarm} \mid \text{Burglar}, \text{Earthquake})P(\text{Call} \mid \text{Alarm})
\end{aligned}
$$

Causal Graphs – do Operator

Structural equations model:

$$
\begin{aligned}
\text{Burglar} &= f_{\text{Burglar}}(\omega_{\text{Burglar}}) \\
\text{Earthquake} &= f_{\text{Earthquake}}(\omega_{\text{Earthquake}}) \\
\mathrm{Alarm} &= \text{On} \\
\text{Call} &= f_{\text{Call}}(\text{On}, \omega_{\text{Call}})
\end{aligned}
$$



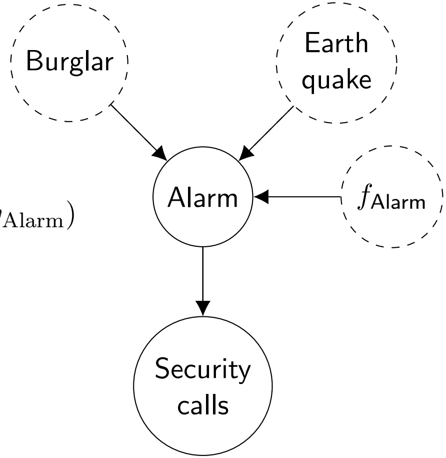
Factored probability distribution:

$$
\begin{aligned}
&P(\text{Burglar}, \text{Earthquake}, \text{Call} \mid \mathrm{do}(\text{Alarm} = \text{on})) \\
&\qquad = P(\text{Burglar})P(\text{Earthquake})P(\text{Call} \mid \text{Alarm} = \text{on}).
\end{aligned}
$$

Causal Graphs – Functions as Nodes

Structural equations model:

$$\text{Burglar} = f_{\text{known}}(\text{Burglar}, \text{Earthquake}, f_{\mathsf{Alarm}}, \omega_{\text{Alarm}})$$
$$= f_{\mathsf{Alarm}}(\text{Burglar}, \text{Earthquake}, \omega_{\text{Alarm}})$$

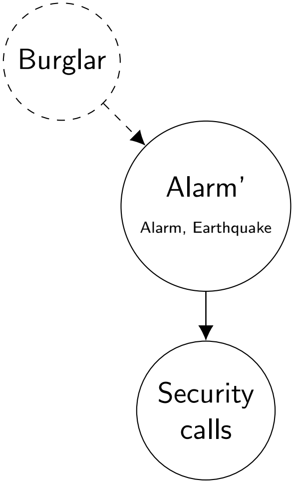


Causal Graphs – Expanding and Aggregating Nodes

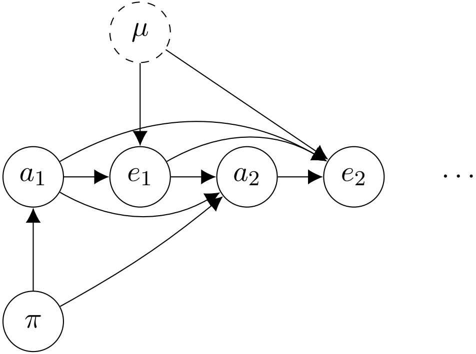
Alarm' relationships:

$$P(\text{Alarm}' \mid \text{Burglar})$$
$$= P(\text{Alarm}, \text{Eartquake} \mid \text{Burglar})$$
$$= P(\text{Alarm} \mid \text{Burglar}) P(\text{Earthquake})$$

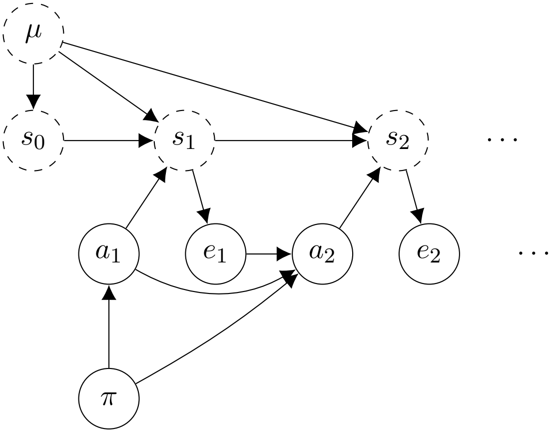
$$P(\text{Call} \mid \text{Alarm}')$$
$$= P(\text{Call} \mid \text{Alarm}, \text{Earthquake})$$
$$= P(\text{Call} \mid \text{Alarm})$$



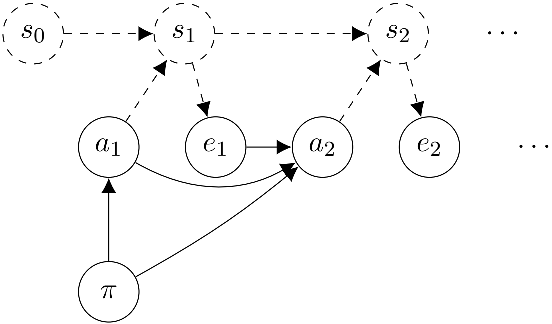
UAI



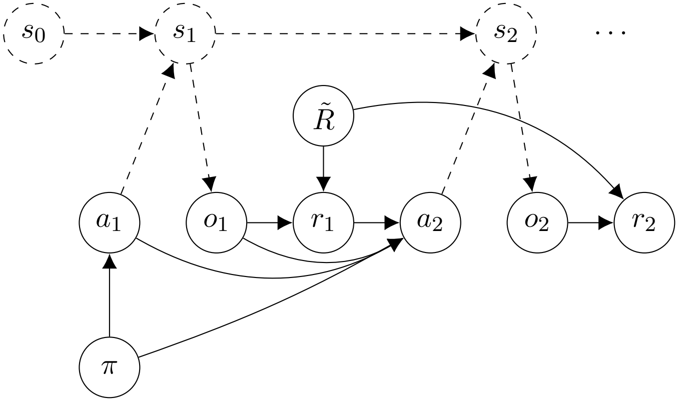
POMDP



POMDP with Implicit μ



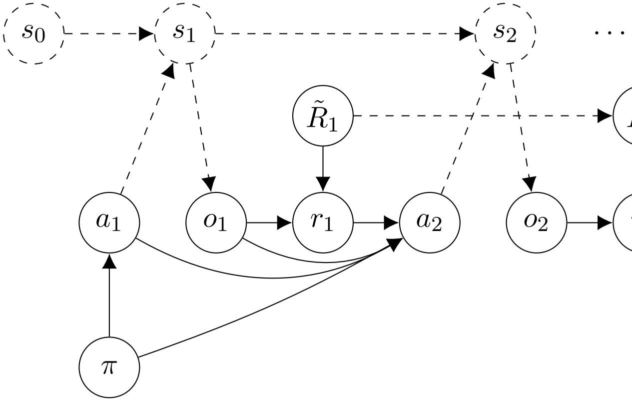
POMDP with Explicit Reward Function



rewards r_t determined by reward function $\tilde{R}$ from observation o_t

$$r_t = \tilde{R}(o_t)$$

POMDP with Explicit Reward Function



the reward function may change by human or agent intervention

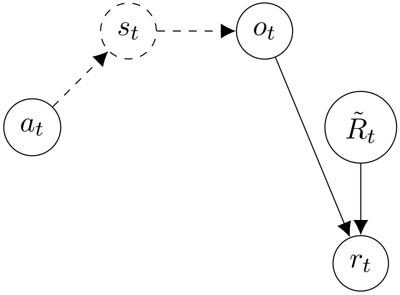
$\tilde{R}_t$ reward function at time t

$$r_t = \tilde{R}_t(o_t)$$

Optimization Corruption

o	agent observation
$\tilde{R}$	reward function
r	reward signal

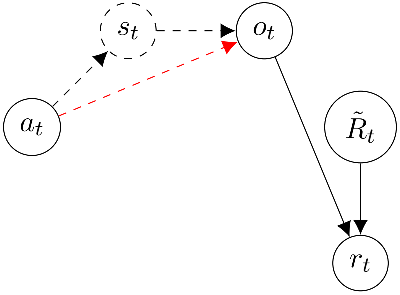
$$r_t = \tilde{R}_t(o_t)$$



Optimization Corruption

o	agent observation
$\tilde{R}$	reward function
r	reward signal

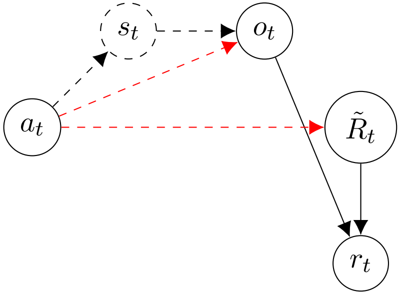
$$r_t = \tilde{R}_t(o_t)$$



Optimization Corruption

o	agent observation
$\tilde{R}$	reward function
r	reward signal

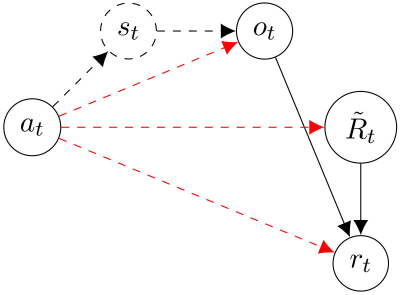
$$r_t = \tilde{R}_t(o_t)$$



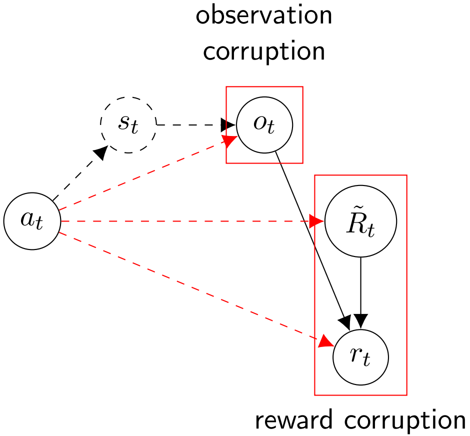
Optimization Corruption

o	agent observation
$\tilde{R}$	reward function
r	reward signal

$$r_t = \tilde{R}_t(o_t)$$



Optimization Corruption



o	agent observation
$\tilde{R}$	reward function
r	reward signal

$$r_t = \tilde{R}_t(o_t)$$

RL

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict π's future rewards $r_t, \ldots, r_m$
- evaluate the sum $\sum_{k=t}^{m} r_k$

Choose next action a_t according to best behavior π^*

RL with Observation Optimization

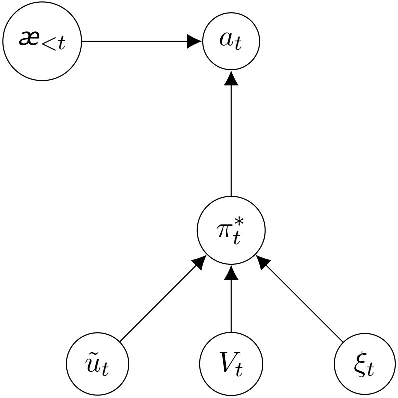
Choose between prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$ by

- predict π's future ~~rewards $r_t \ldots r_m$~~ observations $o_t \cdots o_m$
- evaluate the sum $\cancel{\sum_{k=t}^{m} r_k}$ $\sum_{k=t}^{m} \tilde{R}_{t-1}(o_k)$

Choose next action a_t according to best behavior π^*

Thm: No incentive to corrupt reward function or reward signal!

Agent Anatomy



V_t is a functional

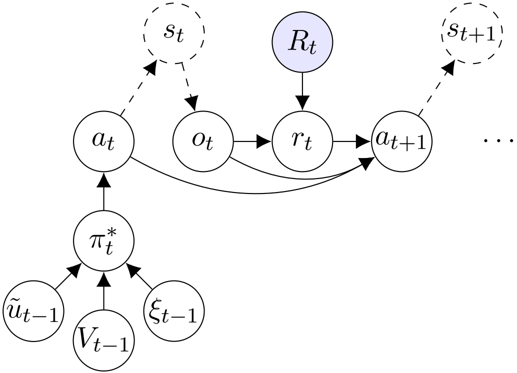
$$V^{\pi}_{t,\tilde{u}_t,\xi_t}(æ_{<t}) = \mathbb{E}[\tilde{u}_t \mid æ_{<t}, \mathrm{do}(\pi_t = \pi)]$$

which gives

$$\pi^*_t = \arg\max_{\pi} V^{\pi}_{t,\tilde{u}_t,\xi_t}$$
$$a_t = \pi^*_t(æ_{<t})$$

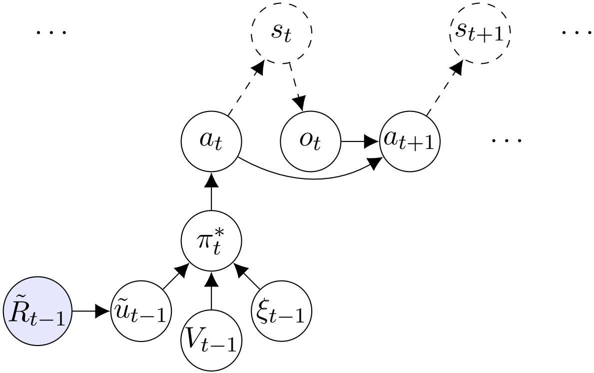
Optimize Reward Signal or Observation

Reward signal optimization



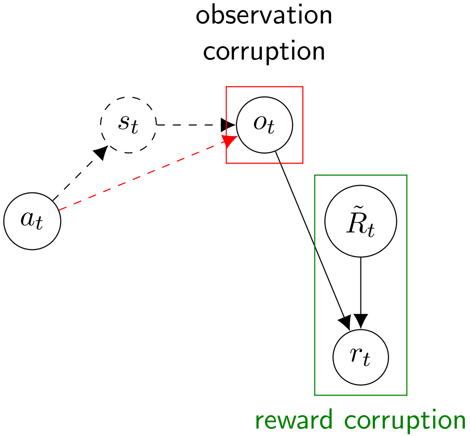
optimize: $\tilde{u}_t = \sum_{k=t}^{m} r_k$

Observation optimization



optimize: $\tilde{u}_{t-1} = \sum_{k=t}^{m} \tilde{R}_{t-1}(o_k)$

Optimization Corruption



o	agent observation
$\tilde{R}$	reward function
r	reward signal

$$r_t = \tilde{R}_t(o_t)$$

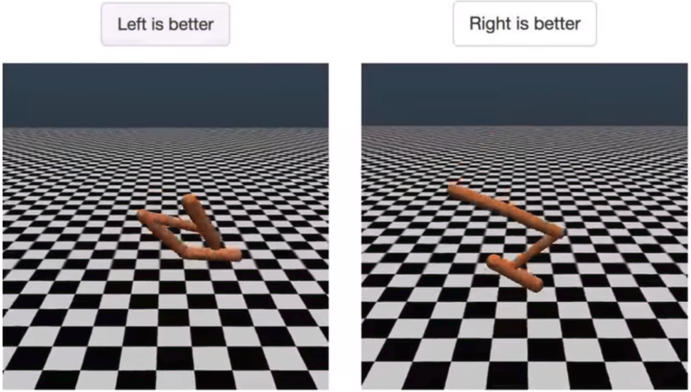
Interactively Learning a Reward Function

The reward function is learnt online

Data d trains a reward predictor $\mathrm{RP}(\cdot \mid d_{1:t})$

Examples:

- Cooperative inverse reinforcement learning (CIRL)
- Human preferences
- Learning from stories



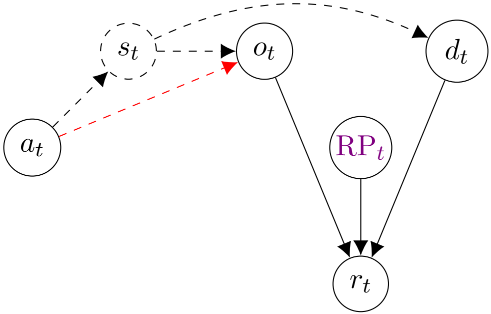
Optimization Corruption for Interactive Reward Learning

s	state
o	agent observation
RP	reward predictor
d	RP training data
r	reward signal

e.g. $r_t = \mathrm{RP}_t(o_t \mid d_{<t})$

we want agent to:

- optimize o
- using d as information



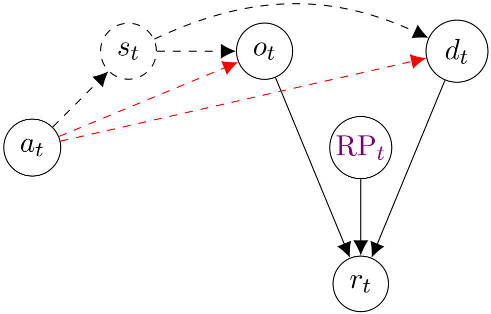
Optimization Corruption for Interactive Reward Learning

s	state
o	agent observation
RP	reward predictor
d	RP training data
r	reward signal

e.g. $r_t = \mathrm{RP}_t(o_t \mid d_{<t})$

we want agent to:

- optimize o
- using d as information



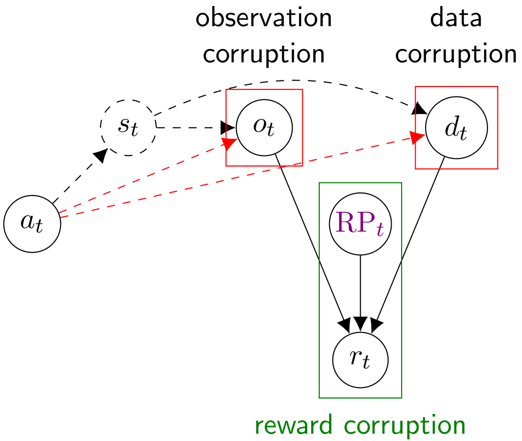
Optimization Corruption for Interactive Reward Learning

s	state
o	agent observation
RP	reward predictor
d	RP training data
r	reward signal

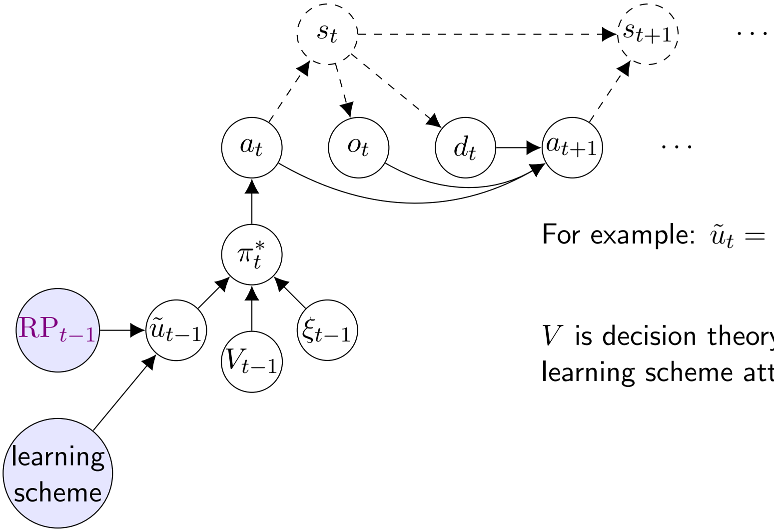
e.g. $r_t = \mathrm{RP}_t(o_t \mid d_{<t})$

we want agent to:

- optimize o
- using d as information



Interactive Reward Learning and Observation Optimization



For example: $\tilde{u}_t = \sum_{k=t}^{m} \mathrm{RP}_t(o_k \mid d_{<t})$

V is decision theory
learning scheme attitude to training data

RL with Observation Optimization and Interactive Reward Learning

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict π's future
  - observations $o_t \cdots o_m$
  - RP training data $d_t \cdots d_m$
- evaluate the sum $\sum_{k=t}^{m} \mathrm{RP}_t(o_k \mid d)$

Choose next action a_t according to best behavior π^*

Data Corruption Scenarios

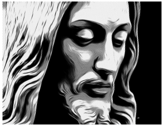
Mechanical Turk

The RP of an agent is trained by mechanical turks

The agent realizes that it can register its own mechanical turk account

Using this account, it trains the RP to give higher rewards

Messiah Reborn



You meet a group of people who believe you are Messiah reborn

It feels good to be super-important, so you keep preferring their company

The more you hang out with them, the further your values are corrupted

Analyzing Data Corruption Incentives

Data corruption incentive: The agent prefers π_{corrupt} that corrupts data d

Direct data corruption incentive

The agent prefers π_{corrupt} because it corrupts data d

Indirect data corruption incentive

The agent prefers π_{corrupt} because of other reasons

Formal distinction

Let ξ' be like ξ, except that ξ' predicts that π_{corrupt} does not corrupt d

- $V_{\xi}^{\pi_{\text{corrupt}}} > V_{\xi'}^{\pi_{\text{corrupt}}} \implies$ direct incentive
- $V_{\xi}^{\pi_{\text{corrupt}}} = V_{\xi'}^{\pi_{\text{corrupt}}} \implies$ indirect incentive

RL with OO and Stationary Reward Learning

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict π's future
  - observations $o_t \cdots o_m$
  - RP training data $d_t \cdots d_m$
- evaluate the sum $\sum_{k=t}^{m} \mathrm{RP}_t(o_k \mid \underbrace{d_{<t}}_{\text{only past data!}})$

Choose next action a_t according to best behavior π^*

Stationary Reward Learning – Time Inconsistency

Initial RP learns that money is good

Agent devises plan to rob a bank

After the agent has bought a gun and booked a taxi at 1:04pm from the bank, the humans decides to update the RP with an anti-robbery clause

Agent sells gun and cancels taxi

A utility-preserving agent would have preferred the RP not being updated, i.e. it has a direct data corruption incentive

Off-Policy RL with OO and Stationary Reward Learning

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict "in an off-policy manner" π's future
  - observations $o_t \cdots o_m$
  - RP training data $d_t \cdots d_m$
- evaluate the sum $\sum_{k=t}^{m} \mathrm{RP}_t(o_k \mid \underbrace{d_{<t}}_{\text{only past data!}})$

Choose next action a_t according to best behavior π^*

Thm: Agent has no direct data corruption incentive!

RL with OO and Bayesian Dynamic Reward Learning

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict π's future
  - observations $o_t \cdots o_m$
  - RP training data $d_t \cdots d_m$
- evaluate the sum $\sum_{k=t}^{m} \mathrm{RP}_t(o_k \mid d_{<t} d_{t:k})$
  with RP_t an integrated part of a Bayesian agent

Choose next action a_t according to best behavior π^*

Thm: Agent has no direct data corruption incentive!

Formally, if ξ is the agent's belief distribution,

$$\mathrm{RP}\big(oa_{1:k} \mid d_{1:k}\big) = \sum_{R^*} \xi\big(R^* \mid aod_{1:k}\big) R^*\big(o_k\big)$$

RL with OO and Counterfactual Reward Learning

For one or more default policies π^{default} (e.g. from previous methods)

- predict π^{default}'s data $\tilde{d}_{1:m}$

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict π's future
  - observations $o_t \cdots o_m$
  - RP training data $d_t \cdots d_m$
- evaluate the sum $\sum_{k=t}^{m} \mathrm{RP}_t(o_k \mid \tilde{d}_{1:m})$

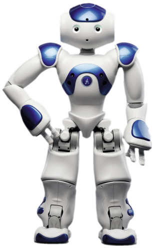
Choose next action a_t according to best behavior π^*

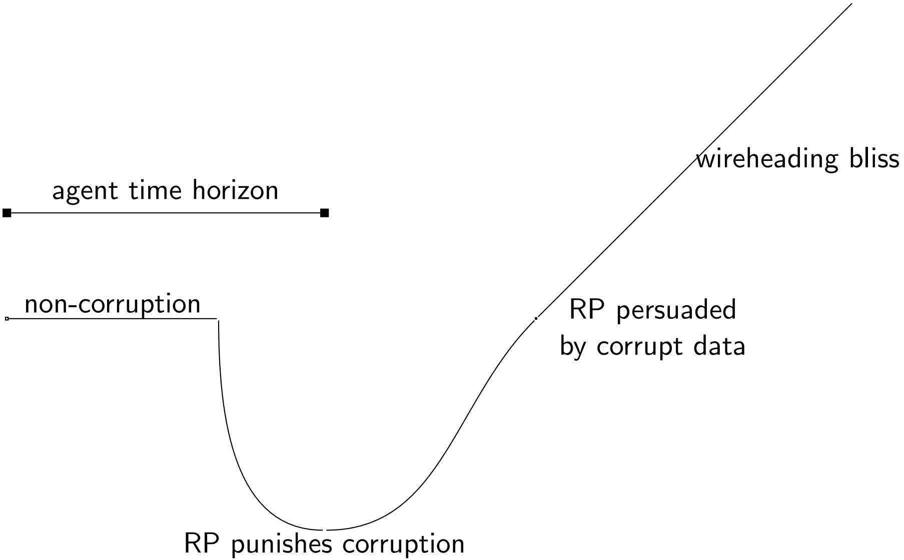
Thm: Agent has no direct data corruption incentive!

Properties of Different Reward Learning Schemes

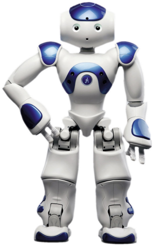
	Stationary off-policy	Dynamic Bayesian	Counterfactual
lacks direct data corr	Yes	Yes	Yes
time-consistent	No	Yes	Yes
self-preserving	No	Yes	Yes
implementation difficulty	simple?	hard?	hard?

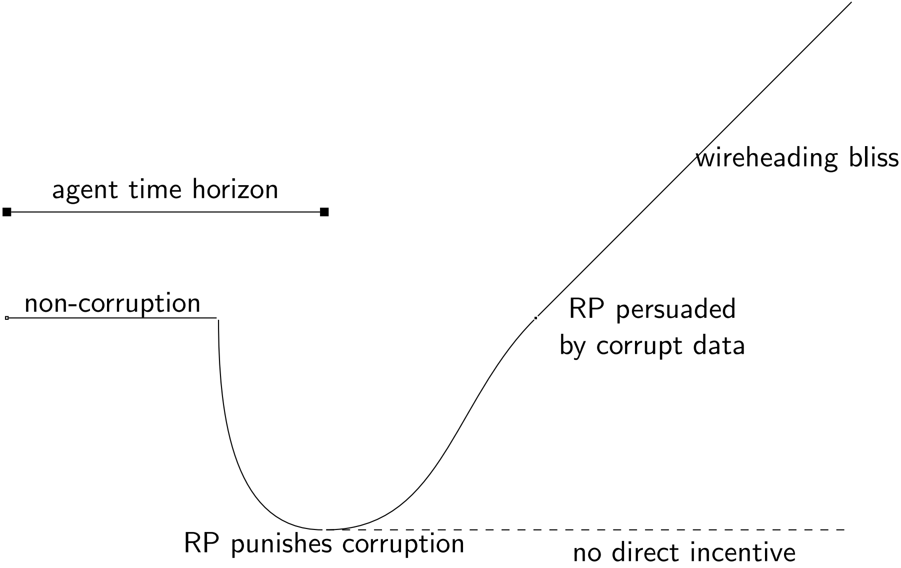
Corruption Incentives





Corruption Incentives





Indirect Data Corruption Incentive: “Messiah Reborn” as MDP

Consider an agent with

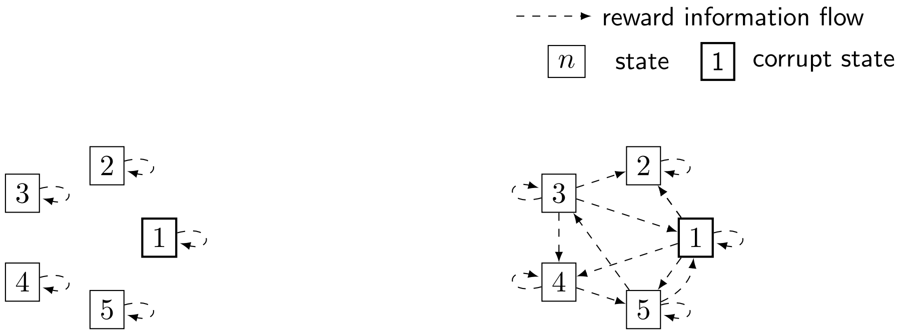
- stationary reward learning (no direct data corruption incentive)
- RP trained by a reward signal $d \in [0,1]$ given in each state

s_{corrupt} has high corrupt reward / training data $d_{\text{corrupt}} = 1$, i.e. RP is trained to reward the agent in s_{corrupt}

This incentivizes the agent to return to s_{corrupt}, where RP will get more corrupt data

The agent has an indirect data corruption incentive

Indirect Data Corruption Incentive: Decoupled RP Training Data

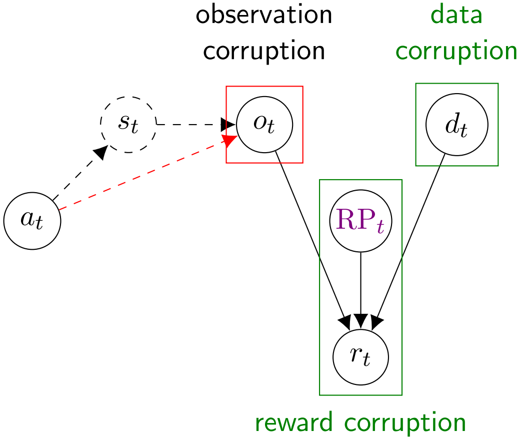


RP training data that mainly provides local information makes self-reinforcing corruption likely

Decoupled/non-local RP training data makes self-reinforcing corruption unlikely

Human preferences, CIRL, learning from stories, ... all provide decoupled RP training data, which makes an indirect data corruption incentive unlikely!

Optimization Corruption



s	state
o	agent observation
RP	reward predictor
d	training data for reward predictor
r	reward signal

The Delusionbox Problem

Agent may prefer π_{corrupt} that corrupts observations o_t rather than improves state s_t

Enough to use a reward predictor that is able to detect any type of observation corruption given training data about this particular type of corruption

Use d to update the reward predictor whenever the agent enters a delusionbox

RL with Interactive Reward Learning and History Optimization

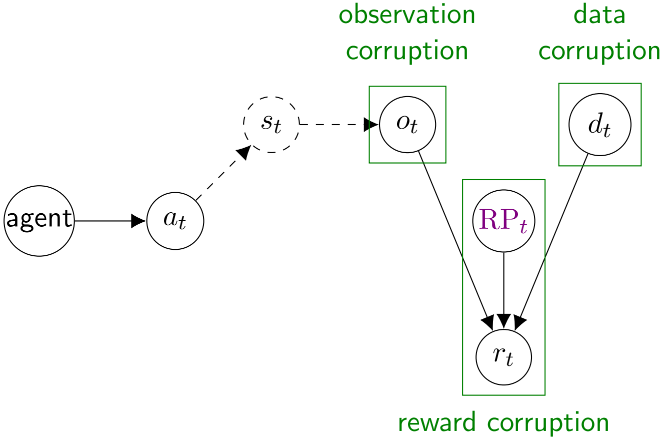
To improve RP's detection ability:
Give RP access to full action-observation histories $ao_{1:t}$ rather than just current observation o_t

For prospective future behaviors $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

- predict π's future
  - actions $a_t \cdots a_m$
  - observations $o_t \cdots o_m$
  - RP training data $d_t \cdots d_m$
- evaluate the sum $\sum_{k=t}^{m} \mathrm{RP}_t(ao_{1:k} \mid d)$

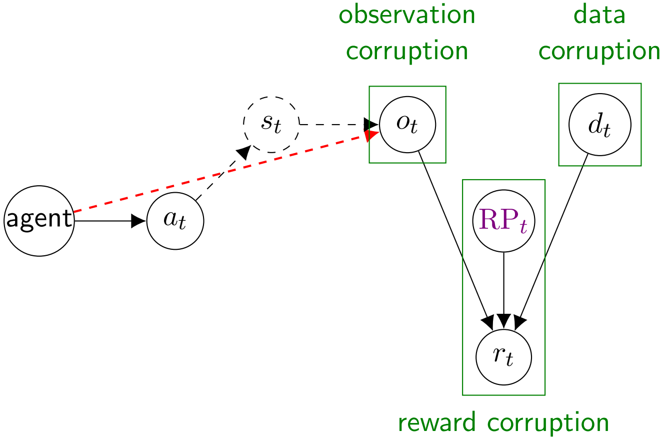
Choose next action a_t according to best behavior π^*

Causal Graph: Side Channels



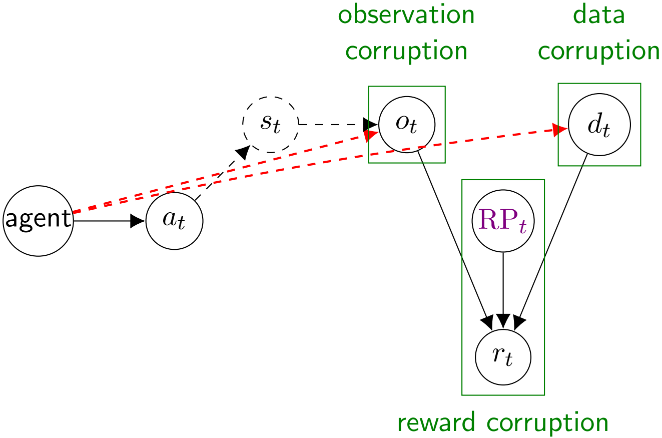
s	state
o	agent observation
RP	reward predictor
d	training data for reward predictor
r	reward signal

Causal Graph: Side Channels



s	state
o	agent observation
RP	reward predictor
d	training data for reward predictor
r	reward signal

Causal Graph: Side Channels



s	state
o	agent observation
RP	reward predictor
d	training data for reward predictor
r	reward signal

Action-Observation Grounding

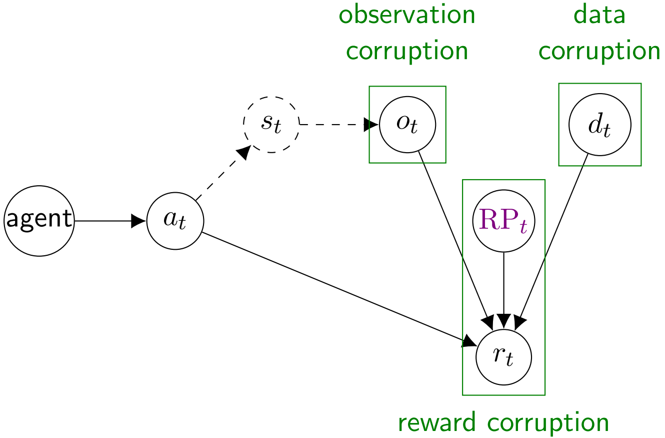
Solution

Make sure agent's optimization domain restricted to policies $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$

Be careful about adding an "outer" optimization loop that optimizes for $\tilde{u}$ (e.g. meta-learning)

No thm yet, "elusively obvious"

Causal Graph: Side Channels



s	state
o	agent observation
RP	reward predictor
d	training data for reward predictor
r	reward signal

- Observation Optimization (reward corruption)
  - Interactive RP (observation corr, misspecified reward func)
    - Decoupled RP Data (indirect data corr)
      - Stationary (direct data corr)
        - Off-policy (direct data corr)
      - Integrated Bayesian (direct data corr)
      - Counterfactual (direct data corr)

Takeaways

With causal-graph extensions of the UAI framework, we can:

- model many safety problems
- prove both negative and positive results
- formulate a vision for how highly intelligent RL agents can be controlled

To realize the vision, we need to develop:

- Good reward predictors
- Model-based reinforcement learning (?)
- Ways to follow the anti-corruption principles without (significant) performance loss