Optimizing Intended Reward Functions

Anca Dragan
Example: Capture influence on human action
Before: not the most efficient
Before: not the most efficient
Before: not the most efficient
Now: sometimes “too” efficient

Planning for autonomous cars that leverage effect on human actions [RSS’16, with Sadigh, Sastry, Seshia]
Add courtesy..

Planning for autonomous cars that leverage effect on human actions [RSS’16, with Sadigh, Sastry, Seshia]
But now, car inches backwards to get you to go!

Planning for autonomous cars that leverage effect on human actions [RSS’16, with Sadigh, Sastry, Seshia]
But now, car inches backwards to get you to go!
comfort
task specification $\rightarrow$ behavior

cost, reward, goal, loss, constraints,

optimization, search, constraint satisfaction, satisficing, RL...
task specification → behavior

optimization, search, constraint satisfaction, satisficing, RL...

cost, reward, goal, loss, constraints, …
What we pretend AI is:

\[ R(s, a) \]

\[ \max \mathbb{E}[\sum_t R(s_t, a_t)] \]
What AI actually is:

\[ R(s, a) \]

\[
\max \mathbb{E}[ \sum_t R(s_t, a_t) ]
\]
AI ≠ optimize specified reward

AI = optimize \textit{intended} reward
Optimize intended reward

\[ R(s, a) \]

\[ \max \mathbb{E} [ \sum_t R(s_t, a_t) ] \]
Optimize intended reward

\[ R_\theta(s, a) \]

parametrization of the reward function

\[ \max \mathbb{E}[\sum_t R_\theta(s_t, a_t)] \]
Optimize intended reward

\[ R_\theta(s, a) \]

\[
\max \mathbb{E} \left[ \sum_t R_\theta(s_t, a_t) \right]
\]

belief over reward parameters

\( b(\theta) \)
Why treat specified rewards as definition?

$R_\theta(s, a)$

$b(\theta) \max \mathbb{E}[\sum_t R_\theta(s_t, a_t)]$

$\tilde{\theta}$
Why treat specified rewards as definition?

\[ R_\theta(s, a) \]

\[ b(\theta) = \begin{cases} 1, & \theta = \tilde{\theta} \\ 0, & \text{else.} \end{cases} \]

\[ \max \mathbb{E}[\sum_t R_\theta(s_t, a_t)] \]
Agents *overlearn* from specified rewards, but *underlearn* from other sources.
Optimize intended reward

\[ R_{\theta}(s, a) \]

max \( \mathbb{E}[ \sum_t R_{\theta}(s_t, a_t)] \)

belief over reward parameters

Oh no what if that breaks?
Humans leak information about the reward.

\[ R_\theta(s, a) \]
How should the robot extract it into an updated belief?
Human feedback, from specifying a reward to turning the robot off, is evidence about the intended reward.

$$R_\theta(s, a)$$

$$\max \mathbb{E}[\sum_t R_\theta(s_t, a_t)]$$

belief over reward parameters

$$b(\theta)$$

don’t step on the carpet
Human feedback, from specifying a reward to turning the robot off, is evidence about the intended reward.
Human feedback, from specifying a reward to turning the robot off, is evidence about the intended reward.

\[ R_{\theta}(s, a) \]

\[ \max \mathbb{E}[\sum R_{\theta}(s_t, a_t)] \]

\[ b'(\theta) \propto b(\theta)P(S_t | \theta) \]

observation (human) model

\( \tilde{\theta} \)

\( \text{don't step on the carpet} \)
What is a human model that can be used to make sense of all these types of human feedback?

\[ R_\theta(s, a) \]

\[ b'(\theta) \propto b(\theta) P(\theta | \theta) \]

- Don't step on the carpet
- \( \tilde{\theta} \)
How can we model reward design/specification as a noisy and suboptimal process?

\[ R_{\theta}(s, a) \]

\[ b'(\theta) \propto b(\theta)P(\tilde{\theta} | \theta) \]

observation (human) model

don’t step on the carpet
Development

score and winning were correlated at development time...

$R_{\tilde{\theta}}$ - score

Deployment

... but no longer correlated at deployment time
We **only** know this about the true reward:

The behavior incentivized by the specified reward in **development** has high true reward.
What you specify is contextualized by the state you specify it in. Robots should interpret it as such.
The behavior incentivized by the specified reward in development has high true reward

\[
P(\tilde{\theta}|\theta^*, M_{devel}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{devel}) \mid \xi \sim P(\xi | \tilde{\theta}, M_{devel})]}
\]
The **behavior incentivized by the specified reward in development** has high true reward

\[ P(\tilde{\theta}|\theta^*, M_{\text{devel}}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{\text{devel}})|\xi \sim P(\xi|\tilde{\theta},M_{\text{devel}})]} \]
The behavior incentivized by the specified reward in development has high true reward

\[ P(\hat{\theta}|\theta^*, M_{devel}) \propto e^{\beta E[R_{\theta^*}(\xi; M_{devel}) | \xi \sim P(\xi | \hat{\theta}, M_{devel})]} \]

Inverse Reward Design
[NIPS’17 with Menell, Milli, Abbeel, Russell]
The behavior incentivized by the specified reward in development has high true reward

\[ P(\tilde{\theta} | \theta^*, M_{\text{devel}}) \propto e^{\beta E[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi | \tilde{\theta}, M_{\text{devel}})]} \]
The behavior incentivized by the specified cost in development has low true cost

\[ P(\tilde{\theta} | \theta^*, M_{\text{devel}}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi | \tilde{\theta}, M_{\text{devel}})]} \]
The behavior incentivized by the specified cost in development has low true cost

\[ P(\tilde{\theta} | \theta^*, M_{\text{devel}}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi | \tilde{\theta}, M_{\text{devel}})]} \]
The behavior incentivized by the specified cost in development has low true cost.

\[ P(\tilde{\theta}|\theta^*, M_{\text{devel}}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi|\tilde{\theta}, M_{\text{devel}})]} \]
The **behavior incentivized by the specified cost in development** has low true cost

\[
P(\tilde{\theta} | \theta^*, M_{\text{devel}}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi | \tilde{\theta}, M_{\text{devel}})]}
\]
The behavior incentivized by the specified cost in development has low true cost

\[ P(\tilde{\theta} | \theta^*, M_{\text{devel}}) \propto e^{\beta E[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi | \tilde{\theta}, M_{\text{devel}})]} \]

<table>
<thead>
<tr>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( \theta_3 )</th>
<th>( \theta_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>maximizing winning</td>
<td>maximizing score</td>
<td>minimizing winning</td>
<td>minimizing score</td>
</tr>
</tbody>
</table>

\( M_{\text{devel}} \)
The behavior incentivized by the specified cost in development has low true cost

\[ P(\tilde{\theta}|\theta^*, M_{\text{devel}}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{\text{devel}}) | \xi \sim P(\xi|\tilde{\theta}, M_{\text{devel}})]} \]
The behavior incentivized by the specified cost in development has low true cost

\[
P(\tilde{\theta}|\theta^*, M_{\text{devel}}) \propto e^{\beta E[R_{\theta^*}(\xi; M_{\text{devel}}) \mid \xi \sim P(\xi|\tilde{\theta}, M_{\text{devel}})]}
\]

\[M_{\text{devel}}\]

\[\theta_1\]
maximizing winning

\[\theta_2\]
maximizing score

\[\theta_3\]
minimizing winning

\[\theta_4\]
minimizing score

\[\checkmark\]
\[\checkmark\]
\[\times\]
\[\times\]
Specified rewards as evidence about the reward

\[ R_\theta(s, a) \]

\[ \tilde{\theta} \]

risk-averse planning

\[
\max_\xi \min_{\theta \in \{\theta_i \sim b_i(\theta)\}} R_\theta(\xi; M_{test}) \\
b'(\theta) \propto b(\theta) P(\tilde{\theta}|\theta)
\]

plan in expectation

\[
\max_\xi \mathbb{E}[R_\theta(\xi; M_{test})|\theta \sim b'(\theta)]
\]
\[ b'(\theta) \propto P(\tilde{\theta}|\theta, M_{\text{train}})b(\theta) \]
Easier, faster, lower regret

This is the independent phase, where you design a desired trajectory separately for 3 environments. The trajectory you want leads to the path in green. The current behavior is shown in the animated images. When you change the slider values, press Recompute Trajectory to show the robot's new path. When you have succeeded in specifying the desired behavior, the trajectory will turn green. Try to specify the correct behavior quickly and in as few recomputations as possible.
Limitations / ongoing work ..
What if we don’t know the important features?!
Inference from raw observations, no direct indicators...

\[ I_s \in \{\text{grass, dirt, target, unk}\} \]
\[ \phi_s \sim \mathcal{N}(\mu_{I_s}, \Sigma_{I_s}) \]
The agent can avoid unintended consequences, even when the features that matter are latent!
Leverage the posterior to identify edge cases

\[ R_\theta(s, a) \]

\[ b'(\theta) \propto b(\theta)P(\tilde{\theta}|\theta) \]

\[ M = \arg \max_m \mathbb{E}_b[H(b) - H(b'|m)] \]
This finds edge-case environments that break the current reward function.
By exposing the designer to these edge cases, regret on held-out environments goes down quickly.
Specified rewards are evidence about the reward.

\[ R_\theta(s, a) \]

\[ b'(\theta) \propto b(\theta) P(\tilde{\theta} | \theta) \]
What is a human model that can be used to make sense of all these types of human feedback?

\[ R_\theta(s, a) \]

Observation (human) model

\[ b'(\theta) \propto b(\theta) P(\text{don't step on the carpet}) | \theta) \]
We know what to do for comparisons.

\[ R_\theta(s, a) \]

\[ b'(\theta) \propto b(\theta)P(\xi|\theta) \]

Active Preference-Based Learning of Reward Functions
[RSS’17 with Sadigh, Seshia, Sastry]
We know what to do for comparisons: model feedback as a reward-rational choice.

\[ R_\theta(s, a) \]

\( b'(\theta) \propto b(\theta) P(\text{observation} | \theta) \)

choices: \( \{ \xi_A, \xi_B \} \)

choose based on reward: \( R_\theta(\xi_A) \) vs \( R_\theta(\xi_B) \)

\[
P(\xi_A | \theta) = \frac{e^{R_\theta(\xi_A)}}{e^{R_\theta(\xi_A)} + e^{R_\theta(\xi_B)}}
\]

don't step on the carpet
We know what to do for demonstrations: model the demo as a reward-rational implicit choice.

$$R_\theta(s, a)$$

$$\tilde{\theta}$$

don’t step on the carpet

$$b'(\theta) \propto b(\theta)P(\text{human} | \theta)$$

choices: $$\{ \xi_i \}$$

choose based on reward: $$R_\theta(\xi_D) \text{ vs } R_\theta(\xi) \forall \xi$$

$$P(\xi_D | \theta) = \frac{e^{R_\theta(\xi_D)}}{\sum_\xi e^{R_\theta(\xi)}}$$

[Ramachandran et al., Bayesian Inverse Reinforcement Learning]
We know what to do for specified rewards

\[ R_{\theta}(s, a) \]

\[ \theta \]

\[ P(\tilde{\theta}|\theta) = \frac{e^{E[R_{\theta}(\xi)|\xi\sim P(\xi|\tilde{\theta}, M_{devel})]}}{\sum_{\theta} e^{E[R_{\theta}(\xi)|\xi\sim P(\xi|\tilde{\theta}, M_{devel})]}} \]
We know what to do for specified rewards: model them as a reward-rational implicit choice.

$$R_\theta(s, a)$$

choices: $$\{ \theta_i \}$$

$$P(\tilde{\theta}|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}}{\sum_{\tilde{\theta}} e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}}$$
We know what to do for specified rewards: model them as a reward-rational implicit choice.

\[ R_{\theta}(s, a) \]

\[ \theta \]

observation (human) model

\[ b'(\theta) \propto b(\theta)P(\theta | \theta) \]

choices:

\[ \{ \theta_i \} \]

choose based on reward:

\[ R_{\theta}(\tilde{\theta})? ! \]

\[ P(\tilde{\theta} | \theta) = \frac{e^{\mathbb{E}[R_{\theta}(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}}{\sum_{\theta} e^{\mathbb{E}[R_{\theta}(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}} \]
We know what to do for specified rewards: model them as a reward-rational implicit choice.

\[ R_\theta(s, a) \]

Observation (human) model

\[ b'(\theta) \propto b(\theta)P(\Theta | \theta) \]

Choices:

\[ \{ \theta_i \} \]

Choose based on reward:

\[ \mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})] \]

Vs

\[ \mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})] \forall \tilde{\theta} \]

\[ P(\tilde{\theta}|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}}{\sum_{\tilde{\theta}} e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}} \]
Reward-rational (implicit) choices

\[ R_\theta (s, a) \]

\[ \theta \]

don't step on the carpet

\[ \mathbb{E}[R_\theta (\xi)|\xi \sim \psi(c^*)] \]

\[ \mathbb{E}[R_\theta (\xi)|\xi \sim \psi(c)] \forall c \]

\[ P(c^*|\theta) = \frac{e^{\mathbb{E}[R_\theta (\xi)|\xi \sim \psi(c^*)]}}{\sum e^{\mathbb{E}[R_\theta (\xi)|\xi \sim \psi(c)]}} \]
How should the robot extract the leaked information into an updated belief?

\[ R_{\theta}(s,a) \]
Key idea: Interpret any type of human feedback as a reward-rational implicit choice.
Human feedback as a reward-rational implicit choice.

\[ R_\theta(s, a) \]

observation (human) model

\[ b'(\theta) \propto b(\theta)P(\theta|\tau) \]

choices: \{ \tau \} (external torques)

choose based on reward:

\[ R_\theta(\xi(\xi_{original}, \tau)) \] (deformed trajectories)

Learning Robot Objectives from Physical Human Interaction
[CoRL'18 with Bajcsy, Losey, O'Malley]
Human feedback as a reward-rational implicit choice.

\[ R_\theta (s, a) \]

\[ b'(\theta) \propto b(\theta) P(\theta | \Theta) \]

Choices:
{press button, do nothing}

Choose based on reward:
\[ R_\theta (\xi_{stopped}) \]
vs
\[ R_\theta (\xi_{planned}) \]

The off-switch game
[IJCAI’17 with Menell, Milli, Abbeel, Russell]
So far, we’ve talked about sources of information that look at human behavior:
To know that you shouldn’t break the vase, you need to see some behavior, e.g.:
What if we don’t see any behavior?

the vase is still here! (and it’s not t=0! )
When the agent is deployed in an environment that the human has been acting in, the state of the environment has information about the human’s intended reward.
The state of the environment as a reward-rational implicit choice.

\[ R_\theta(s, a) \]

Preferences implicit in the state of the world

[ICLR’19 with Shah, Krashenninikov, Alexander, Abbeel]
What reward function is the state consistent with?
What reward function is the state consistent with?

want to break the vase
What reward function is the state consistent with?

don’t care about the vase
What reward function is the state consistent with?

want to not break the vase
Side effects: Room with vase
Desirable side effects: Batteries
AI ≠ optimize specified reward

AI = optimize intended reward
Human feedback as reward-rational implicit choice

$R_\theta(s, a)$

choices: \[ \{ c \} \]

choose based on reward:

\[
\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c^*)] \]

vs

\[
\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c)] \forall c
\]

\[
P(c^* | \theta) = \frac{e^{\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c^*)]}}{\sum e^{\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c)]}}
\]
Agents overlearn from specified rewards, but leave other information on the table.

We can read the right amount of information into each source by interpreting them as reward-rational implicit choices.
Agents overlearn from specified rewards, but leave other information on the table.

We can read the right amount of information into each source by interpreting them as reward-rational implicit choices.
task specification $\rightarrow$ behavior

cost, reward, loss, constraints,…

optimization, search, constraint satisfaction, satisficing, RL…
\[ \max \mathbb{E}\left[ \sum_t R_{\tilde{\theta}}(s_t, a_t) \right] \]
Assistance Games

\[ \max \mathbb{E}[\sum_t R_\theta(s_t, a^R_t, a^H_t)] \]
Thanks!