Environment Design for Inverse Reinforcement Learning

Thomas Kleine Buening[1] * Victor Villin[2] * Christos Dimitrakakis[2 3]

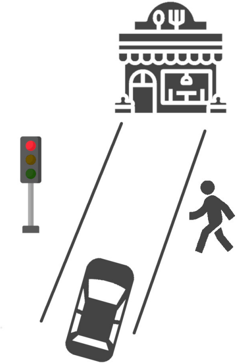
[1]The Alan Turing Institute, London, UK
[2]Université de Neuchâtel, Neuchâtel, Switzerland
[3]University of Oslo, Oslo, Norway

The Alan Turing Institute

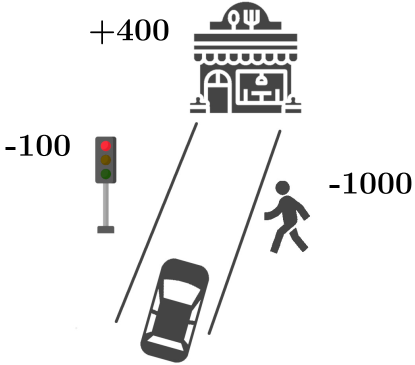
unine Université de Neuchâtel

UNIVERSITY OF OSLO

Inverse Reinforcement Learning (IRL)

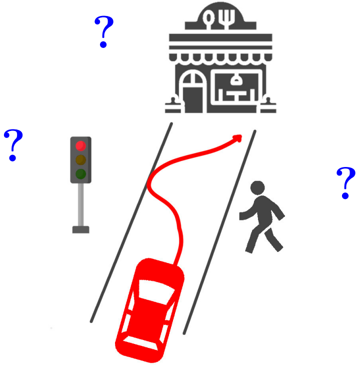


Inverse Reinforcement Learning (IRL)



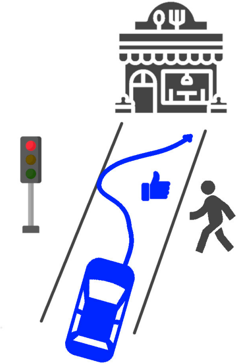
- Engineering suitable reward functions **is hard**

Inverse Reinforcement Learning (IRL)



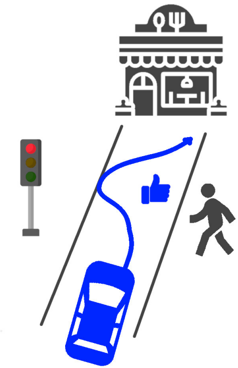
- Engineering suitable reward functions **is hard**
- **Goal**: retrieve the rewards motivating the expert

Inverse Reinforcement Learning (IRL)



- Engineering suitable reward functions **is hard**
- **Goal**: retrieve the rewards motivating the expert
- Useful to train autonomous agents
- Ensures AI alignment

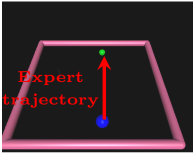
Inverse Reinforcement Learning (IRL)



- Engineering suitable reward functions **is hard**
- **Goal**: retrieve the rewards motivating the expert
- Useful to train autonomous agents
- Ensures AI alignment
- $\neq$ Imitation learning

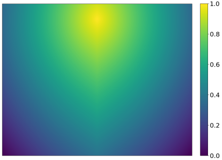
Challenges in IRL

Challenges in IRL



Demonstrations

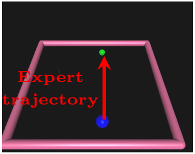
True rewards

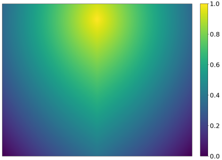


Plausible estimates

(1) **Non-uniqueness** of estimates even with $t \to \infty$

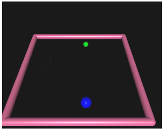
Challenges in IRL



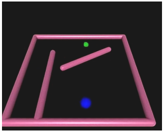


Demonstrations | **True rewards** | Plausible estimates

(1) **Non-uniqueness** of estimates even with $t \to \infty$



$\xrightarrow{\text{infer}} \bar{R} \xrightarrow{?}$

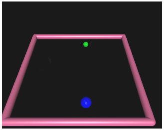


(2) Infer **robust** estimates of the rewards

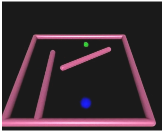
Challenges in IRL

Demonstrations | **True rewards** | Plausible estimates

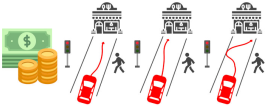
(1) **Non-uniqueness** of estimates even with $t \to \infty$



$\xrightarrow{\text{infer}} \bar{R} \xrightarrow{?}$

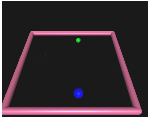


(2) Infer **robust** estimates of the rewards



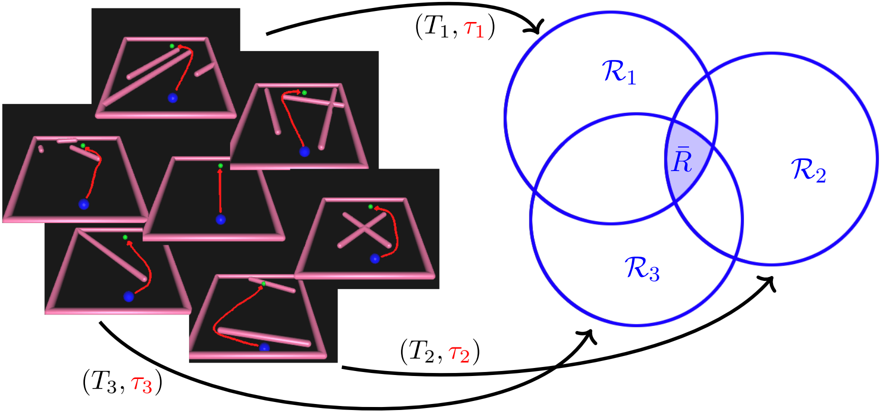
(3) Sample **Efficiency**

Reward Inference from **Multiple Environments**



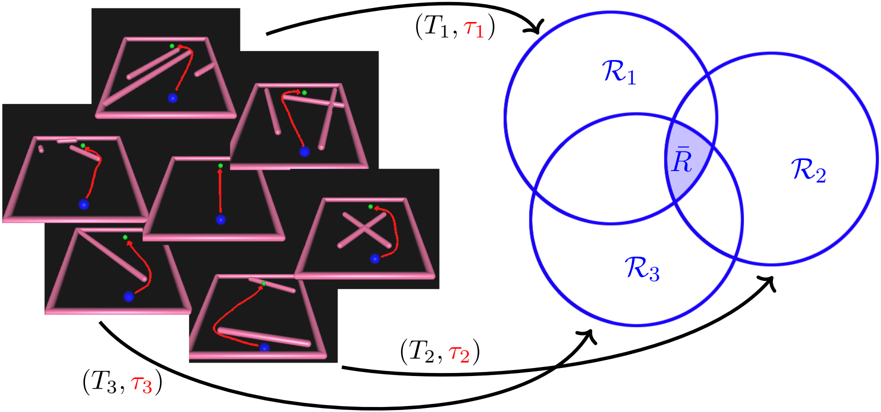
Reward Inference from **Multiple Environments**

- Infer what **commonly** motivates the expert in each environment



Reward Inference from **Multiple Environments**

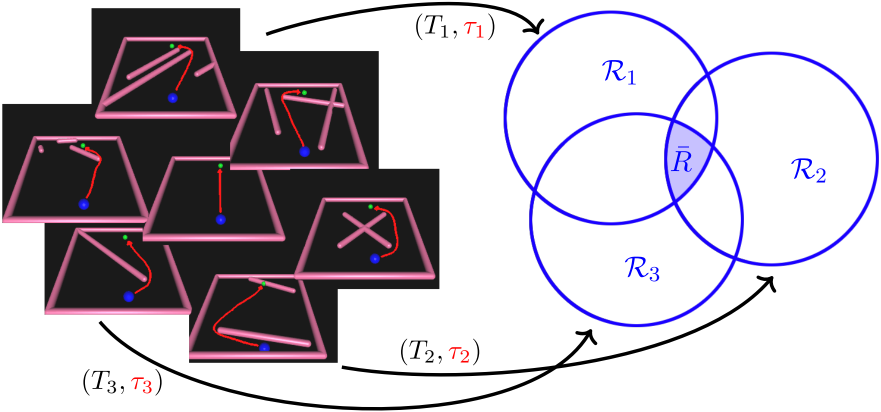
- Infer what **commonly** motivates the expert in each environment



- We **Extend** IRL to **multiple environments** (Bayesian IRL, AIRL)

Reward Inference from **Multiple Environments**

- Infer what **commonly** motivates the expert in each environment



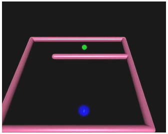
- We **Extend** IRL to **multiple environments** (Bayesian IRL, AIRL)

Problem: We are **limited** in human (expert) data budget!
Which environments should we **choose** ?

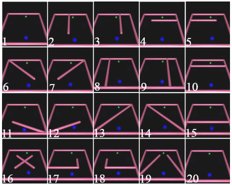
Environment design

Framework:

(1) Select environment T



choose $\longleftarrow$

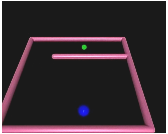


Environment set $\mathcal{T}$

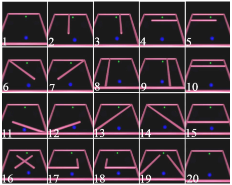
Environment design

Framework:

(1) Select environment T

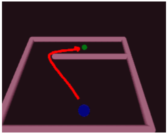


choose
$\longleftarrow$



Environment set $\mathcal{T}$

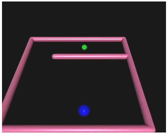
(2) Observe trajectories τ



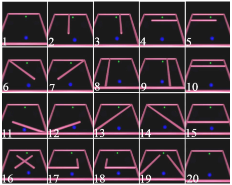
Environment design

Framework:

(1) Select environment T

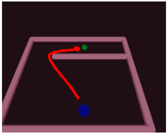


choose $\longleftarrow$



Environment set $\mathcal{T}$

(2) Observe trajectories τ

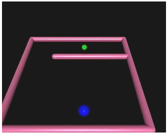


(3) Update our beliefs $\mathbb{P}$ about the true rewards (**Multiple Environment IRL**)

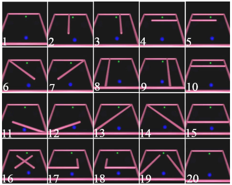
Environment design

Framework:

(1) Select environment T

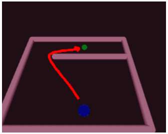


choose $\longleftarrow$



Environment set $\mathcal{T}$

(2) Observe trajectories τ



(3) Update our beliefs $\mathbb{P}$ about the true rewards (**Multiple Environment IRL**)

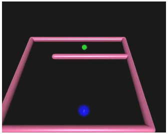
(4) Repeat

Environment design

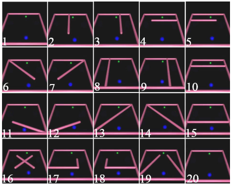
Framework:

$\arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$

(1) Select environment T

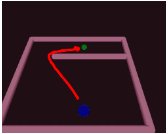


choose $\leftarrow$



Environment set $\mathcal{T}$

(2) Observe trajectories τ



(3) Update our beliefs $\mathbb{P}$ about the true rewards (**Multiple Environment IRL**)

(4) Repeat

Minimax Regret for Environment Design

- Value $\mathcal{V}(T, R, \pi)$ for each environment-reward-policy tuple T, R, π.
- Belief $\mathbb{P}(R)$ over reward function candidates.

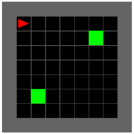
$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

- We pick the **worst-case** environment

$$\boxed{T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)}$$

- An environment with **large regret** is an environment we **have a lot to learn from** potentially.

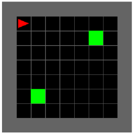
$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$
$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$



$T_0(\gamma = 0.9)$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_{R} \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_{T} \min_{\pi} \text{Regret}(T, \mathbb{P}, \pi)$$



$T_0(\gamma = 0.9)$

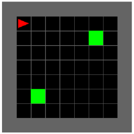
R_1

R_2

$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

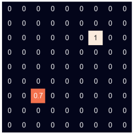


$T_0(\gamma = 0.9)$

R_1

R_2

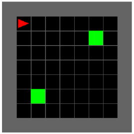
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$



$\bar{R}$ ('best guess')

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

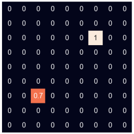


$T_0(\gamma = 0.9)$

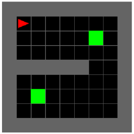
R_1

R_2

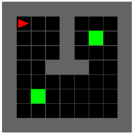
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



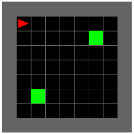
T_1



T_2

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_{R} \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_{T} \min_{\pi} \text{Regret}(T, \mathbb{P}, \pi)$$

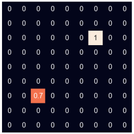


$T_0(\gamma = 0.9)$

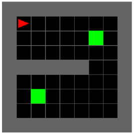
R_1

R_2

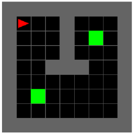
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



T_1

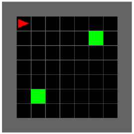


T_2

$$\arg\min_{\pi} \text{Regret}(T, \mathbb{P}, \pi) = \arg\max_{\pi} \mathcal{V}(T, \bar{R}, \pi)$$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

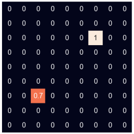


$T_0(\gamma = 0.9)$

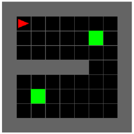
R_1

R_2

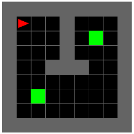
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



T_1



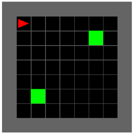
T_2

$$\arg\min_\pi \text{Regret}(T, \mathbb{P}, \pi) = \arg\max_\pi \mathcal{V}(T, \bar{R}, \pi)$$

$$\min_\pi \text{Regret}(T_1, \mathbb{P}, \pi) = 0.3 \times [\gamma^6 - \gamma^6]$$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

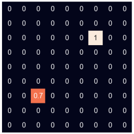


$T_0(\gamma = 0.9)$

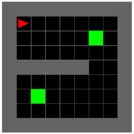
R_1

R_2

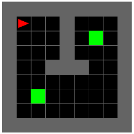
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



T_1



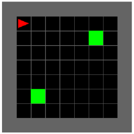
T_2

$$\arg\min_\pi \text{Regret}(T, \mathbb{P}, \pi) = \arg\max_\pi \mathcal{V}(T, \bar{R}, \pi)$$

$$\min_\pi \text{Regret}(T_1, \mathbb{P}, \pi) = 0.3 \times [\gamma^6 - \gamma^6] + 0.7 \times [\gamma^6 - \gamma^6] = 0$$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

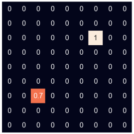


$T_0(\gamma = 0.9)$

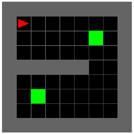
R_1

R_2

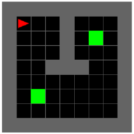
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R\sim\mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



T_1



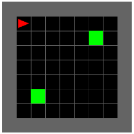
T_2

$$\arg\min_\pi \text{Regret}(T, \mathbb{P}, \pi) = \arg\max_\pi \mathcal{V}(T, \bar{R}, \pi)$$

$$\min_\pi \text{Regret}(T_1, \mathbb{P}, \pi) = 0.3 \times [\gamma^6 - \gamma^6] + 0.7 \times [\gamma^6 - \gamma^6] = 0$$
$$\min_\pi \text{Regret}(T_2, \mathbb{P}, \pi) = 0.3 \times [\gamma^{12} - 0]$$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$

$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

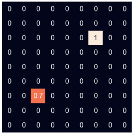


$T_0(\gamma = 0.9)$

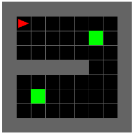
R_1

R_2

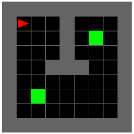
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R \sim \mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



T_1



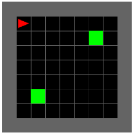
T_2

$$\arg\min_\pi \text{Regret}(T, \mathbb{P}, \pi) = \arg\max_\pi \mathcal{V}(T, \bar{R}, \pi)$$

$$\min_\pi \text{Regret}(T_1, \mathbb{P}, \pi) = 0.3 \times [\gamma^6 - \gamma^6] + 0.7 \times [\gamma^6 - \gamma^6] = 0$$

$$\min_\pi \text{Regret}(T_2, \mathbb{P}, \pi) = 0.3 \times [\gamma^{12} - 0] + 0.7 \times [\gamma^6 - \gamma^6] = 0.06$$

$$\text{Regret}(T, \mathbb{P}, \pi) = \sum_R \mathbb{P}(R)[\mathcal{V}^*(T, R) - \mathcal{V}(T, R, \pi)]$$
$$T^* = \arg\max_T \min_\pi \text{Regret}(T, \mathbb{P}, \pi)$$

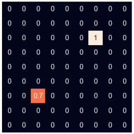


$T_0(\gamma = 0.9)$

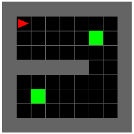
R_1

R_2

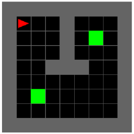
$$\mathbb{P}(R_1) = 0.3$$
$$\mathbb{P}(R_2) = 0.7$$
$$\bar{R} = \mathbb{E}_{R\sim\mathbb{P}}[R]$$



$\bar{R}$ ('best guess')



T_1



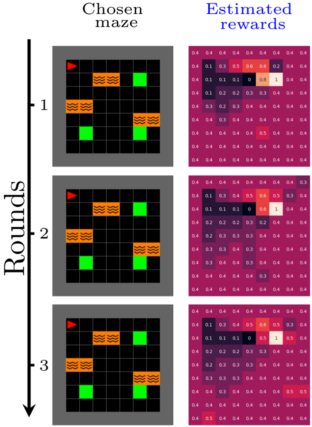
T_2

$$\arg\min_\pi \text{Regret}(T, \mathbb{P}, \pi) = \arg\max_\pi \mathcal{V}(T, \bar{R}, \pi)$$

$$\min_\pi \text{Regret}(T_1, \mathbb{P}, \pi) = 0.3 \times [\gamma^6 - \gamma^6] + 0.7 \times [\gamma^6 - \gamma^6] = 0$$
$$\min_\pi \text{Regret}(T_2, \mathbb{P}, \pi) = 0.3 \times [\gamma^{12} - 0] + 0.7 \times [\gamma^6 - \gamma^6] = 0.06$$

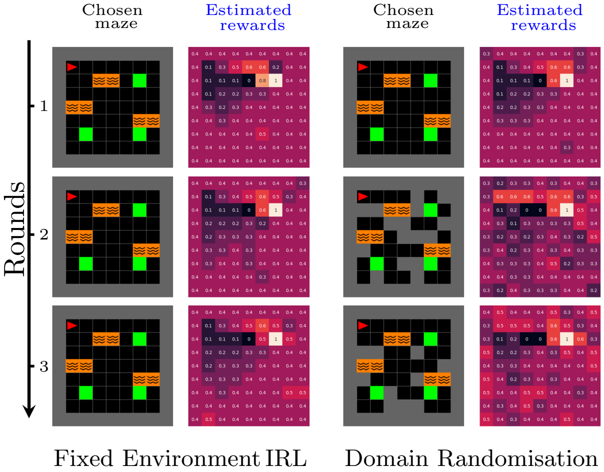
We end up choosing T_2

Can we recover **performance**-relevant aspects of the **true** reward function ?

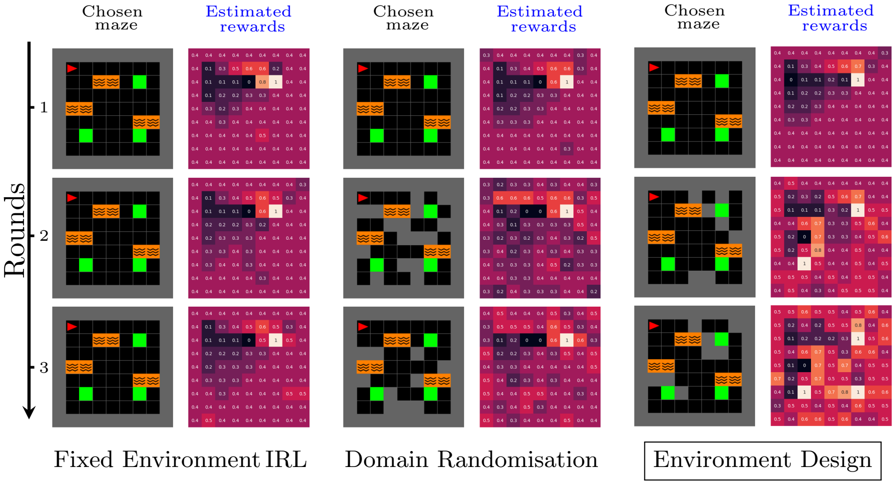


Fixed Environment IRL

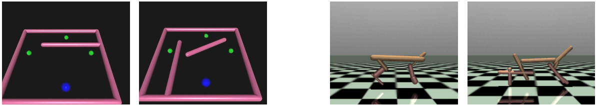
Can we recover **performance**-relevant aspects of the **true** reward function ?



Can we recover **performance**-relevant aspects of the **true** reward function ?



Environment design **robustifies** estimates

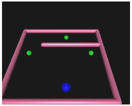


Demo Test Demo Test

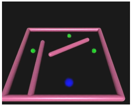
Continuous maze HalfCheetah

Examples of demo and test environments.

Environment design **robustifies** estimates

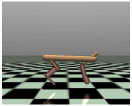


Demo

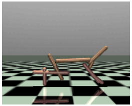


Test

Continuous maze



Demo



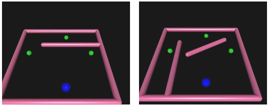
Test

HalfCheetah

Examples of demo and test environments.

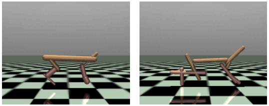
	Continuous maze		Hopper		HalfCheetah		Swimmer	
	Demo	Test	Demo	Test	Demo	Test	Demo	Test
ED-AIRL	**68±04**	**71±02**	**63±07**	52±04	**40±11**	35±13	80±19	69±12
DR-AIRL	52±07	53±12	59±06	**56±04**	**40±13**	**40±11**	45±04	53±05
AIRL	33±09	52±07	38±03	34±04	29±09	16±07	40±09	44±08

Environment design **robustifies** estimates



Demo Test

Continuous maze



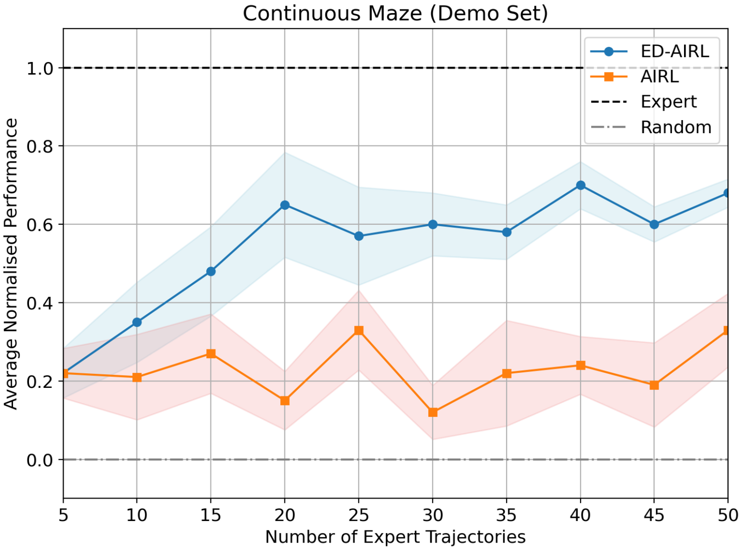
Demo Test

HalfCheetah

Examples of demo and test environments.

	Continuous maze		Hopper		HalfCheetah		Swimmer	
	Demo	Test	Demo	Test	Demo	Test	Demo	Test
ED-AIRL	**68±04**	**71±02**	**63±07**	52±04	**40±11**	35±13	80±19	69±12
DR-AIRL	52±07	53±12	59±06	**56±04**	**40±13**	**40±11**	45±04	53±05
AIRL	33±09	52±07	38±03	34±04	29±09	16±07	40±09	44±08
Imitation learning								
DR-RIME	-105±12	-52±03	61±01	53±02	-21±08	-11±09	-05±01	-04±01
GAIL	20±05	17±01	40±02	34±01	-12±02	-06±01	111±00	110±01
BC	11±00	22±00	-12±02	-06±01	-23±01	-14±01	**124±01**	**130±01**

Environment design **unlocks** IRL's **sample efficiency**



Conclusion (Environment Design for IRL)

The Alan Turing Institute

unine Université de Neuchâtel

UNIVERSITY OF OSLO

Conclusion (Environment Design for IRL)

- Extension of IRL inference methods to **multiple environments**

The Alan Turing Institute

unine
Université de Neuchâtel

UNIVERSITY OF OSLO

Conclusion (Environment Design for IRL)

- Extension of IRL inference methods to **multiple environments**
- **Automated** environment design choosing **worst-case** environments

The Alan Turing Institute

unine Université de Neuchâtel

UNIVERSITY OF OSLO

Conclusion (Environment Design for IRL)

- Extension of IRL inference methods to **multiple environments**
- **Automated** environment design choosing **worst-case** environments
- **Superior** sample efficiency

The Alan Turing Institute

unine Université de Neuchâtel

UNIVERSITY OF OSLO

Conclusion (Environment Design for IRL)

- Extension of IRL inference methods to **multiple environments**
- **Automated** environment design choosing **worst-case** environments
- **Superior** sample efficiency
- Recovers **most** performance-relevant aspects of **unknown** reward functions

The Alan Turing Institute

unine Université de Neuchâtel

UNIVERSITY OF OSLO

Conclusion (Environment Design for IRL)

- Extension of IRL inference methods to **multiple environments**
- **Automated** environment design choosing **worst-case** environments
- **Superior** sample efficiency
- Recovers **most** performance-relevant aspects of **unknown** reward functions
- Estimates rewards that **transfer better** to **new** transition dynamics

The Alan Turing Institute

unine Université de Neuchâtel

UNIVERSITY OF OSLO