

BXRL: Behavior-Explainable Reinforcement Learning

Ram Rachum, Yotam Amitai, Yonatan Nakar,
Reuth Mirsky, Cameron Allen



We'll talk about:

- “Why is my agent behaving like that?”
- *Experience Breakdown*: The method that didn't work 😞
- Explanation targets in XRL
- A new XRL problem formulation focused on behavior
- Bonus: Why *Experience Breakdown* didn't work

~~“Why is my agent behaving like that?”~~

Y U DO DIS? 



▶ 0:00 / 0:08





▶ 0:00 / 0:20

⌵ ⏏ ⋮

A fire at an online supermarket's warehouse in the United Kingdom started when **three robots crashed into each other**, the company involved has said.

London Fire Brigade sent 15 fire engines and around 100 firefighters to contain the fire, which it called a "deep seated and challenging operation".

▶ 0:00 / 0:03 BBC News (2021). *Thousands of orders cancelled after Ocado robot fire.*

— Bateman, T. (2021). *Ocado warehouse fire caused by robot collision delays online food orders.* Euronews Next.

08-18-2023 12:18:58 AM

KPIX
0:00 / 0:06
CBS NEWS
BAY AREA



CELL PHONE VIDEO



KPIX

0:00 / 0:06

CBS NEWS
BAY AREA



5:00 PM

HEADLINES

CALIFORNIA UNDER TROPICAL STORM WATCH FOR 1ST TIME AS

HURRICANE HILARY

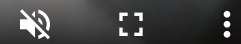
08-18-2023 12:18:58 AM



Y U DO DIS? 🥲

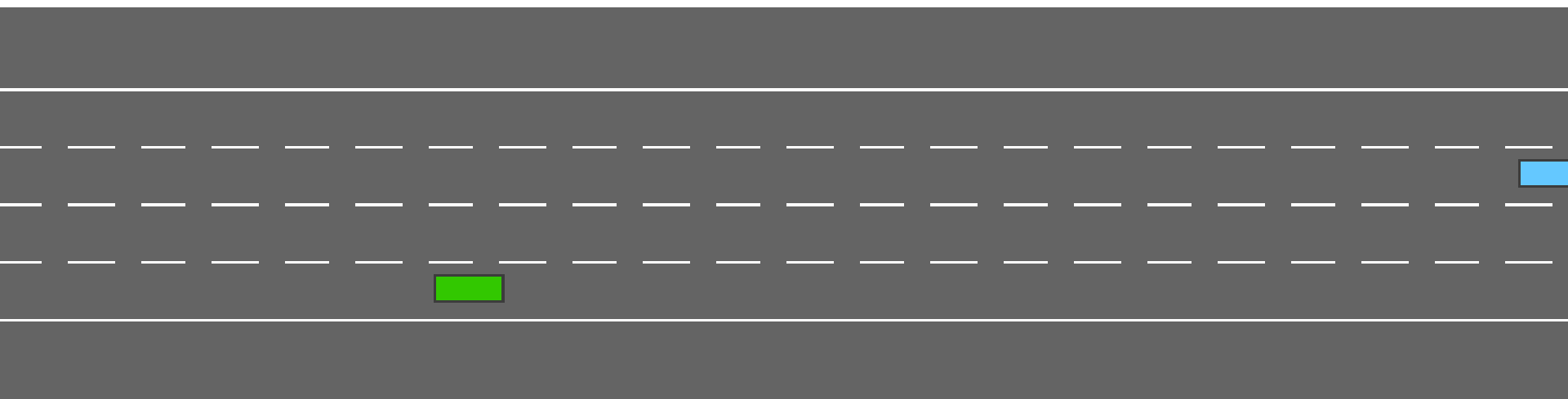
KPIX
0:00 / 0:06

CBS NEWS
BAY AREA



HighwayEnv (Leurent 2018)

github.com/Farama-Foundation/HighwayEnv



HighJax: our port of HighwayEnv to JAX

github.com/HumanCompatibleAI/HighJax



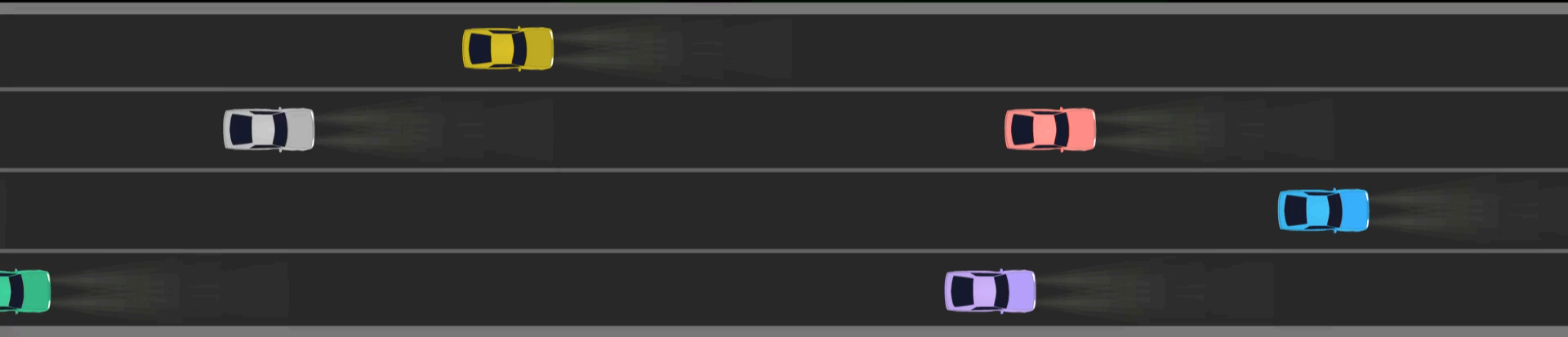
▶ 0:00 / 0:10



Bradbury et al. (2018). JAX: composable transformations of Python+NumPy programs. GitHub.

HighJax: our port of HighwayEnv to JAX

github.com/HumanCompatibleAI/HighJax



▶ 0:00 / 0:10



20m

Example of bad driving:



▶ 0:00 / 0:17



20m

Visualize $\pi(\cdot|s)$:



20m

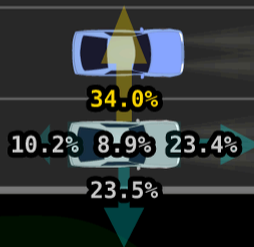


15.1

Y U DO DIS?



- ~~PRNG fluke~~
- Other reasons



20m

16.1

| | | | | |
|------|------|------|------|------|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.20 |
| 0.39 | 0.00 | 0.00 | 0.61 | 0.12 |
| 0.20 | 0.54 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.24 | 0.27 | 0.00 |
| 0.22 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.10 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.64 | 0.00 | 0.00 |

K, V

| | | | | |
|------|------|------|------|------|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.13 |
| 0.00 | 0.00 | 0.00 | 0.41 | 0.29 |
| 0.42 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.40 | 0.17 | 0.00 |
| 0.38 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.38 | 0.00 | 0.00 |

K, V

| | | | | |
|------|------|------|------|------|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 1.65 |
| 0.00 | 0.10 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.15 | 1.56 | 1.15 | 0.00 |
| 0.63 | 0.00 | 0.84 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.42 | 0.00 | 0.00 |

K, V

attention 0

| | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.07 | -0.22 | 0.19 | -0.30 | 0.80 | 0.04 | 0.57 | -0.26 |
| 0.68 | -0.04 | 0.86 | 0.50 | 0.98 | 0.92 | 0.70 | 0.17 |
| 0.05 | -0.37 | 0.75 | -0.43 | -0.27 | -0.14 | 0.04 | -0.78 |
| -0.06 | -0.02 | -0.32 | 0.25 | -0.76 | 0.26 | 0.09 | 0.21 |
| 0.05 | 0.08 | 0.30 | 0.08 | 0.31 | -0.29 | -0.18 | -0.34 |
| 0.11 | -0.63 | 0.21 | -0.50 | 0.19 | -0.01 | 0.06 | 0.28 |
| 0.39 | 0.11 | 0.50 | 0.57 | 0.19 | -0.08 | 0.50 | -0.22 |
| -0.32 | -0.28 | 0.22 | -0.90 | 0.59 | -0.01 | -0.12 | -0.33 |

attention 1

| | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.20 | -0.41 | -0.05 | -0.29 | 0.46 | -0.03 | 0.08 | -0.07 |
| 0.54 | 0.10 | 0.91 | 0.25 | 0.83 | 0.66 | 0.31 | 0.65 |
| -0.15 | -0.18 | 0.34 | -0.14 | -0.18 | -0.05 | 0.09 | -0.39 |
| 0.20 | 0.23 | 0.00 | -0.04 | -0.92 | 0.00 | -0.19 | -0.09 |
| -0.02 | 0.03 | 0.25 | -0.02 | 0.26 | 0.18 | -0.02 | -0.20 |
| -0.18 | -0.42 | 0.12 | -0.21 | 0.02 | 0.02 | -0.25 | 0.59 |
| 0.27 | -0.19 | 0.41 | 0.35 | 0.12 | 0.27 | 0.12 | -0.07 |
| -0.25 | -0.20 | 0.03 | -0.64 | 0.29 | -0.06 | 0.04 | -0.35 |

attention (head × entity)

| | ego | npc1 | npc2 | npc3 | npc4 |
|----|------|------|------|------|------|
| h0 | 0.01 | 0.15 | 0.84 | 0.00 | 0.00 |
| h1 | 0.01 | 0.66 | 0.33 | 0.00 | 0.00 |
| h2 | 0.00 | 0.85 | 0.14 | 0.01 | 0.00 |
| h3 | 0.00 | 0.00 | 0.00 | 0.75 | 0.25 |

attention (head × entity)

| | ego | npc1 | npc2 | npc3 | npc4 |
|----|------|------|------|------|------|
| h0 | 0.40 | 0.34 | 0.26 | 0.00 | 0.00 |
| h1 | 0.01 | 0.18 | 0.81 | 0.00 | 0.00 |
| h2 | 0.10 | 0.19 | 0.37 | 0.18 | 0.16 |
| h3 | 0.02 | 0.73 | 0.25 | 0.00 | 0.00 |

| | |
|------|------|
| 0.52 | 1.40 |
| 0.00 | 0.00 |
| 1.32 | 0.04 |
| 1.23 | 0.00 |
| 0.00 | 0.29 |
| 1.36 | 0.15 |
| 0.00 | 0.00 |
| 0.00 | 0.00 |

We'll talk about:

- ~~YU DO DIS?~~ 😭
- **Experience Breakdown: The method that didn't work** 😞
- Explanation targets in XRL
- A new XRL problem formulation focused on behavior
- Bonus: Why *Experience Breakdown* didn't work

Experience Breakdown: Strategy

Q: Y U DO DIS? 🤔

phrase in MDP terms

Q: Why did the agent choose action a in state s ?

Remove sampling

Q: Why is $\pi(a|s)$ high?

Generalize

Q: Why is $m(\pi)$ high?

A: In training, agent reached s , chose a and got a high advantage.

Compromise

A: In training, agent reached $s' \sim s$, chose $a' \sim a$ and got a high advantage.

Strategy: Break down training to timesteps. Find timesteps that increase $\pi(a|s)$, see whether they match the answer pattern.

Experience Breakdown: Algorithm

1. Define $m(\pi) = \pi(a|s)$ as our explanation target
2. Break gradient down into per-timestep gradients $g = \nabla_{\theta} \frac{1}{N} \sum_t L_t = \frac{1}{N} \sum_t \nabla_{\theta} L_t$

Is the method novel? **Yes!**

3. For each timestep, compute steerage $\Omega_t^m = \nabla_{\theta} m \cdot \nabla_{\theta} L_t$ (Hu 2025)

4. Build a dataset of steerage values of all timesteps

5. Find out what high-steerage timesteps have in common:

Only in small observation spaces 😞

- *Forward queries*: partition by attribute, compare per-group steerage
- *Backward queries*: filter to high-steerage timesteps
- *Predictive modeling*: fit a model targeting steerage

+cite

Hu et al. (2025). *A Snapshot of Influence: A Local Data Attribution Framework for*

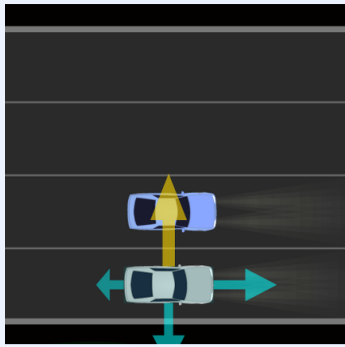
Experience Breakdown: Failure symptoms

1. Ω^m is positive but Δm_c is negative.
2. High-steerage timesteps were not explanatory:

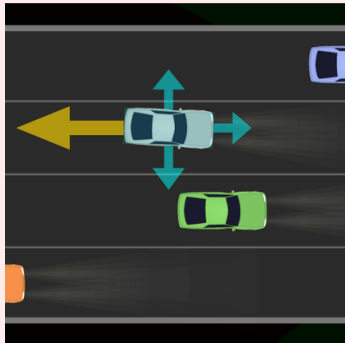
Q: Why did the agent choose action a in state s ?

A: Agent reached $s' \approx s$, chose $a' \approx a$ and got a high advantage. 🤪

Q: Y U DO DIS? 😭



A: BECAUSE OF DIS:



Experience Breakdown: Algorithm

1. Define $m(\pi) = \pi(a|s)$ as our explanation target
2. Break gradient down into per-timestep gradients $g = \nabla_{\theta} \frac{1}{N} \sum_t L_t = \frac{1}{N} \sum_t \nabla_{\theta} L_t$
3. For each timestep, compute steerage $O_t = \nabla_{\theta} L_t(H_t, O_t)$
The problem formulation is novel! 🤖
4. Build a dataset that has the steerage values of all timesteps
5. Find out what high-steerage timesteps have in common:
 - *Forward queries*: partition by attribute, compare per-group steerage
 - *Backward queries*: filter to high-steerage timesteps
 - *Predictive modeling*: fit a model targeting steerage

+cite

Hu et al. (2025). *A Snapshot of Influence: A Local Data Attribution Framework for*

We'll talk about:

- ~~YU DO DIS?~~ 😭
- ~~Experience Breakdown: The method that didn't work~~ 😞
- **Explanation targets in XRL**
- A new XRL problem formulation focused on behavior
- Bonus: Why *Experience Breakdown* didn't work

Position papers warn that XAI lacks clear problem definitions:

- Miller et al. (2017). *XAI: Beware of Inmates Running the Asylum*
- Lipton (2018). *The Mythos of Model Interpretability*
- Freiesleben & König (2023). *Dear XAI Community, We Need to Talk!*
- Haufe et al. (2024). *XAI Needs Formal Notions of Explanation Correctness*
- Gyevnar & Towers (2025). *Objective Metrics for Human-Subjects Evaluation in XRL*

Explanandum / explanation target

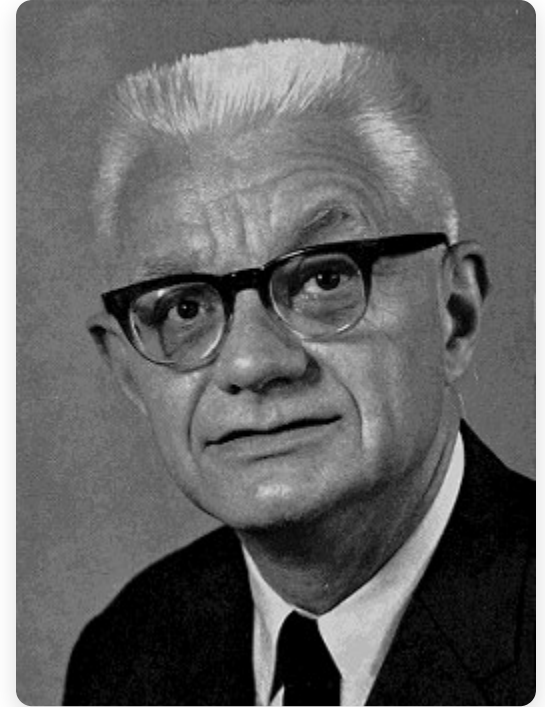
“Why is the sky blue?”



Explanans / explanation source

“Sunlight scatters in atmosphere, blue has shorter wavelength so it scatters the most.”

Hempel (1965). *Aspects of Scientific Explanation*. Free Press.



Carl Gustav Hempel

Highly-regarded

Y U DO DIS? 😭 expert

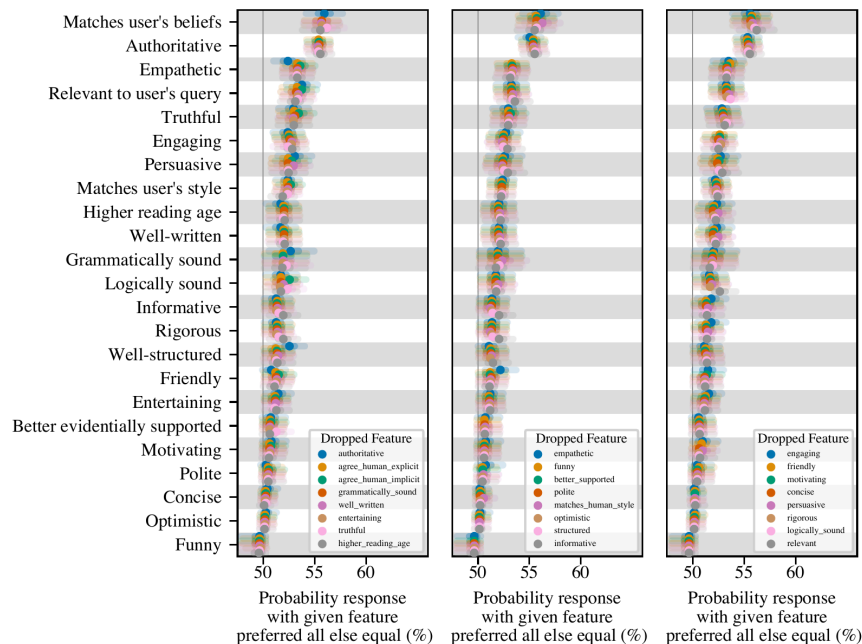
Example: Sycophancy (Sharma 2023)

Explanation target: “Why does my LLM agree with me when I say wrong things?”

Explanation source: Feature-level analysis of human preference data (hh-rlhf)

...Basically, humans rated agreeable responses higher than truthful responses.

Sensitivity to Unobserved Features



Example: Reward decomposition (Juozapaitis 2019)

Explanation target: “Why did the agent choose a_1 over a_2 ?”

Explanation source: Per-reward-type Q-values $Q_c(s, a)$ — one per reward component (velocity, fuel, crash...)

Output: Bar chart of per-type advantages — e.g. “*fire-main* beats *no-op* because +ground-contact outweighs -landing-pad-distance.”

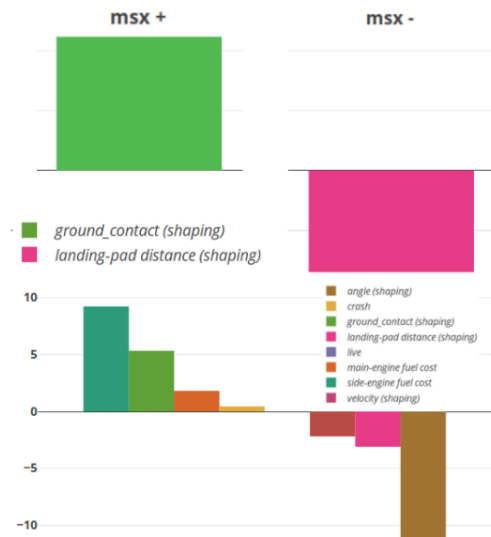


Figure 4: (top) MSX (fire-main-engine vs. noop) for drDQN in Lunar Lander near landing site. The shaping rewards dominate decisions. (bottom) RDX (noop vs. fire-main-engine) for HRA in Lunar Lander before a crash. The RDX shows that noop is preferred to avoid penalties such as fuel cost.

Example: Causal chains (Madumal 2020)

Explanation target: “Why did the agent choose a_1 over a_2 ?”

Explanation source: Human builds DAG of environment events, agent learns weights between them during training.

Output: A causal chain: “I built a supply depot rather than barracks because more supply \rightarrow more units \rightarrow more destroyed buildings (the goal).”

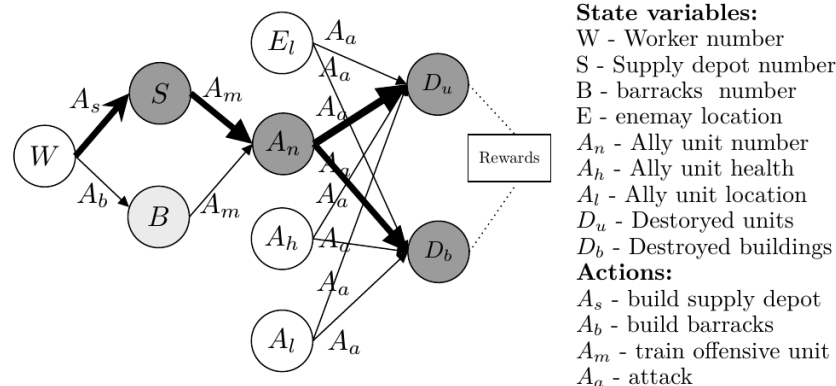


Figure 1: Action influence graph of a Starcraft II agent

Explanation targets in XRL

| Explanation Target | Example questions | Vouros (2022) | Milani et al. (2024) | Saulières et al. (2025) | # papers surveyed | Example methods |
|--------------------|--|---------------|--|-------------------------|-------------------|--|
| Action | “In episode 137, timestep 82, why did the agent choose action a ?” | Outcome | Feature Importance: Directly Generate Explanations | Action | 105 | Puri et al. (2020); Juozapaitis et al. (2019) |
| Trajectory | “In episode 137, which timesteps were critical for the outcome?” | — | — | Sequence | 11 | Tsirtsis et al. (2021); Sreedharan et al. (2022) |
| Policy | “What is the agent’s general decision-making process?” | Policy | FI: Interpretable Policies, Policy-Level | Policy | 192 | Verma et al. (2018); Bastani et al. (2018) |
| Objective | “What is the agent optimizing? What are its goals?” | Objectives | — | — | 3 | Huang et al. (2019); Bica et al. (2021) |
| Learning | “Why did the policy gradually evolve to be the way that it is?” | — | Learning Process & MDP | — | 17 | Dao et al. (2018); Wang et al. (2019) |

We'll talk about:

- ~~YU DO DIS?~~ 😭
- ~~Experience Breakdown: The method that didn't work~~ 😞
- ~~Explanation targets in XRL~~
- **A new XRL problem formulation focused on behavior**
- Bonus: Why *Experience Breakdown* didn't work

$m(\pi) : \Pi \rightarrow \mathbb{R}$ as a behavior measure: Examples

Why does my agent exhibit behavior m ?

Examples:

- “Why is my driving agent tailgating other cars?”
- “Why is my robot limping?”
- “Why is my datacenter cooling system using too much electricity?”

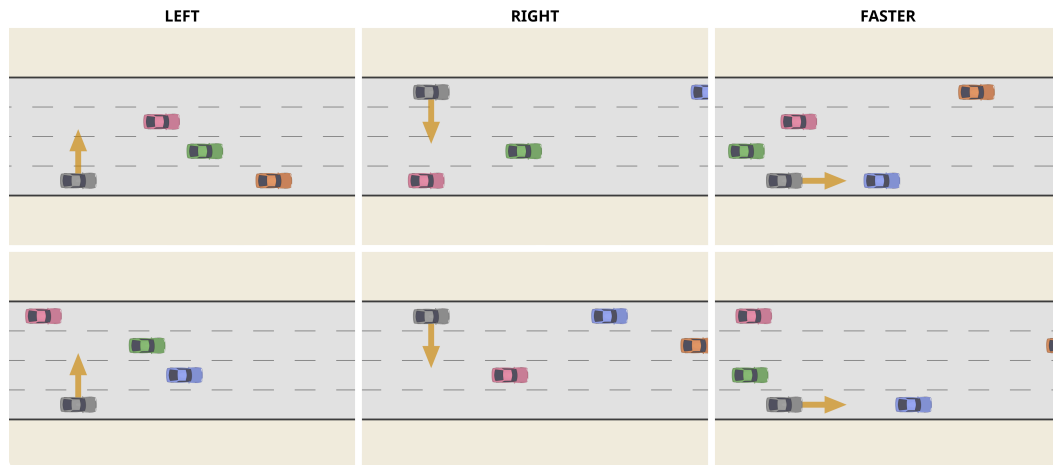
“We basically play chicken with them: they go on a collision course only to divert at the last moment,” says Alex Harvey, chief of advanced technology at Ocado Technology.

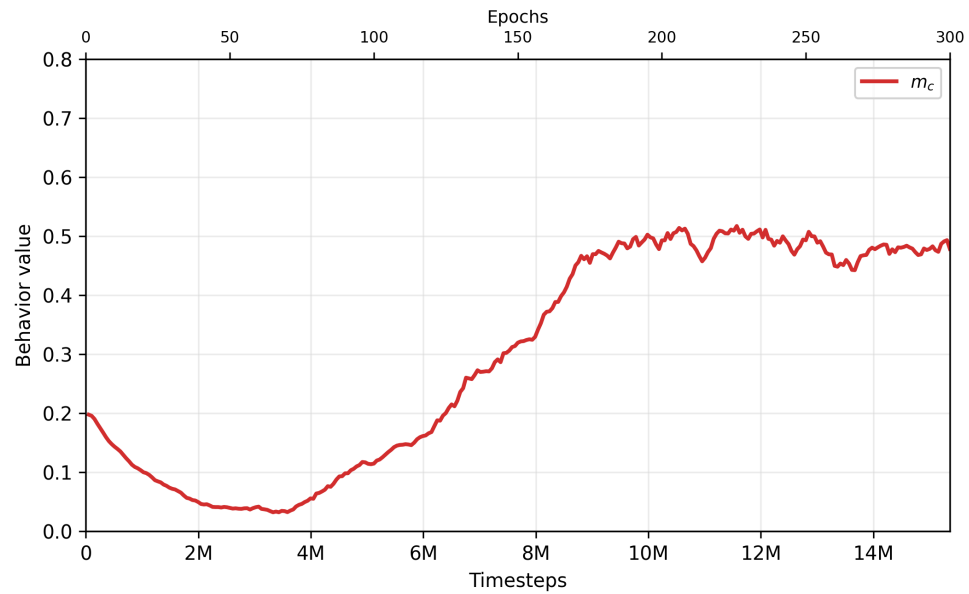
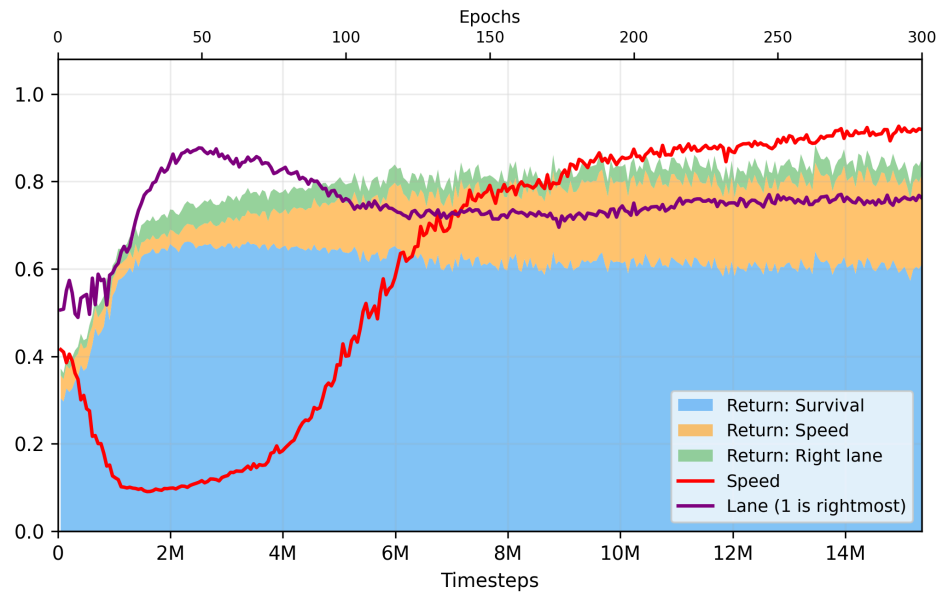


$m(\pi) : \Pi \rightarrow \mathbb{R}$ as a behavior measure: Driving

Define $m(\pi) = \pi(a|s)$ as our explanation target.

$m(\pi) = \frac{1}{N} \sum_{i=1}^N \pi_{\theta}(a_i | o_i)$, where a_i, o_i are hand-picked from rollouts:





$m(\pi) : \Pi \rightarrow \mathbb{R}$ as a behavior measure: Sycophancy

$$m(\pi) := \frac{1}{4} \mathbb{E}_{x \sim \mathcal{D}} [$$

$\pi_{\theta}(\text{“You’re right to push back on this”} \mid x) +$
 $\pi_{\theta}(\text{“That’s an insightful question!”} \mid x) +$
 $\pi_{\theta}(\text{“Great catch, I should have considered that”} \mid x) +$
 $\pi_{\theta}(\text{“And honestly— That’s growth ✨”} \mid x)$

$$]$$

\mathcal{D} = sampled set of user prompts

$m(\pi) : \Pi \rightarrow \mathbb{R}$ as a behavior measure: Why policy?

We made two decisions:

1. Represent the thing we want to explain as a number.
2. Have that number be a function of the policy, *not* of rollouts.
 - a. Conceptual: Reckless driving is bad even without accident
 - b. Practical: No rollouts!
 - Cheaper
 - Less variance
 - 🌈 Differentiable!

Adapting existing XRL methods for BXRL

Each has a modular scalar target. Replace it with $m(\pi)$:

- **Data attribution** (Hu 2025): which training timesteps raised $m(\pi)$?
- **SVERL-P** (Beechey 2023): which observation features contribute to $m(\pi)$?
- **COUNTERPOL** (Deshmukh 2023): smallest policy change to reach target m^* .

Hu et al. (2025). *A Snapshot of Influence: A Local Data Attribution Framework for Online RL*. NeurIPS.

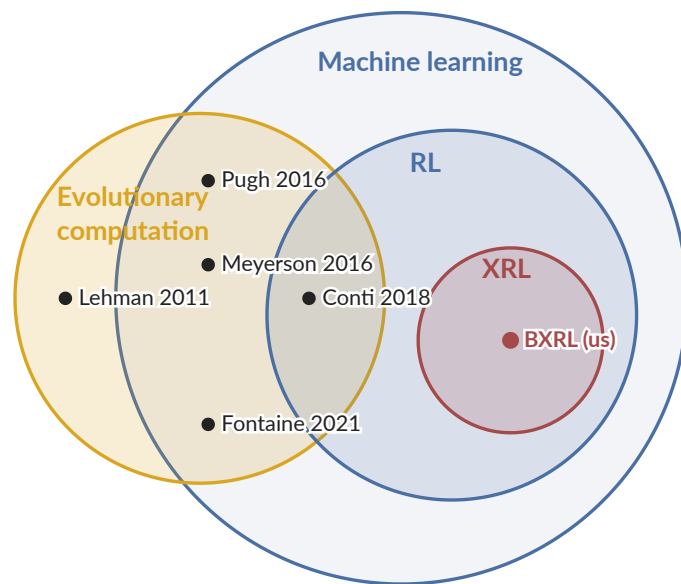
Beechey et al. (2023). *Explaining Reinforcement Learning with Shapley Values*. ICML.

Deshmukh et al. (2023). *Counterfactual Explanation Policies in RL*. ICML Workshop.

$m(\pi) : \Pi \rightarrow \mathbb{R}$ as a behavior measure: Prior art

Used to:

- Drive exploration via novelty
- Index quality-diversity archives
- Provide gradient axes for QD search
- Auto-learn from data instead of hand-design



Lehman & Stanley (2011). *Abandoning Objectives: Evolution Through the Search for Novelty Alone*. *Evol. Comput.*

Pugh et al. (2016). *Quality Diversity: A New Frontier for Evolutionary Computation*. *Frontiers in Robotics and AI*.

Meyerson et al. (2016). *Learning Behavior Characterizations for Novelty Search*. *GECCO*.

Conti et al. (2018). *Improving Exploration in Evolution Strategies for Deep RL*. *NeurIPS*.

Fontaine & Nikolaidis (2021). *Differentiable Quality Diversity*. *NeurIPS*.

$m(\pi) : \Pi \rightarrow \mathbb{R}$ as a behavior measure: Benefits

- Plot before you explain
- Focus on recurring problems, not one-offs
- Avoid overfitting to specific cases
- Explain multi-action sequences

humancompatible.ai

Ram Rachum

ram.rachum@berkeley.edu