ICML 2026 · DEMO WORKSHOP · ORAL PRESENTATION

MOLTEN POT

Evaluations & Datasets for Social Offline Reinforcement Learning

Juan Claude Formanek · Marcel Hedman · Kale-ab Tessera · Christopher C. Holmes · Jonathan P. Shock

SOCIAL DISTRIBUTION SHIFT

The offline RL field has mainly focused on addressing **state and action** distribution shift — where the learned policy drifts from the behaviour policy that generated the data. Offline MARL extends this to multi-agent settings: benchmarks like **OG-MARL** study learning **coordination** from offline, multi-agent data — but only **shared-reward**, with no mixed motives. An overlooked form of shift for real deployment is **Social Distribution Shift** — where a co-player's motives change, and the agent must infer this and adjust its policy accordingly.

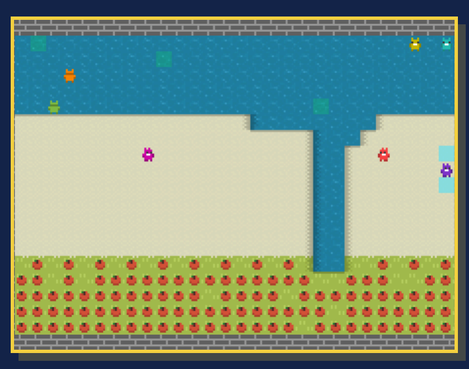
DATA REGIME \ SOCIAL STRUCTURE	Cooperative	Mixed-motive
Online	SMAC, MA MuJoCo, PettingZoo	Melting Pot, JaxMARL
Offline	OG-MARL	**MOLTEN POT** offline · mixed-motive

THE SOCIAL ENVIRONMENTS

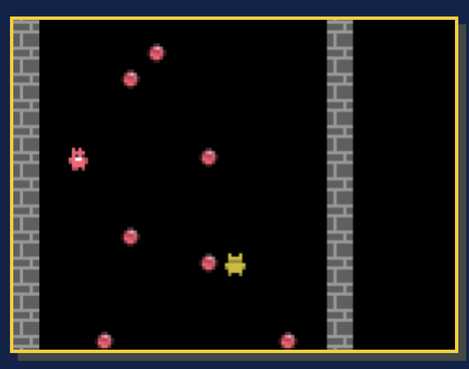
An offline benchmark on five **mixed-motive Melting Pot** substrates. Each **scenario** pairs a substrate with a different population of co-players — agents that may cooperate, exploit, or reciprocate — and a **focal** agent must do well inside it.

5 SUBSTRATES · **47** SCENARIOS · **3** SETTINGS · **1 TB** DATA

FIVE MIXED-MOTIVE WORLDS

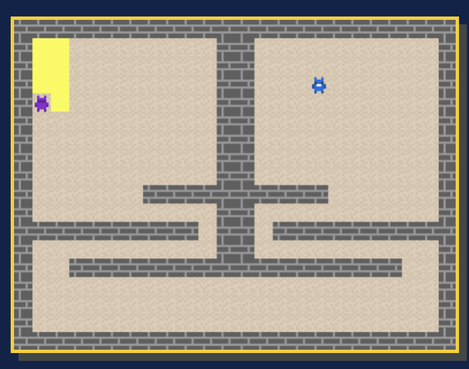


Clean Up — public goods

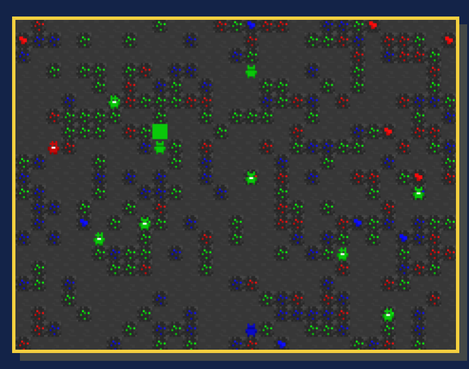


Coins — reciprocity

Coop Mining — trust



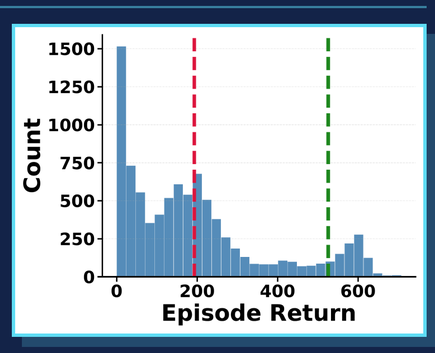
Commons Harvest — commons



Allelopathic — convention

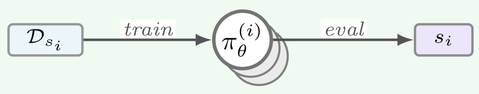
SKILL-MIXED DATA

Each scenario's dataset comes from an independent PPO run: we log trajectories sampled **uniformly across training** — from the near-random policy at the start to the fully converged policy at the end. That makes every dataset deliberately **skill-mixed**, spanning the whole range of behaviour quality rather than only expert demonstrations. We normalise episode return so **0 = a random policy** and **1 = the dataset's own p_{90}**, its best-performing band — making scores comparable across all 47 scenarios.



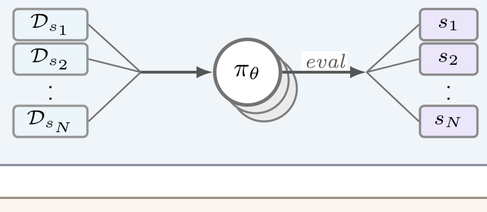
THREE SETTINGS

Evaluation Setting 1: Single-Scenario Offline Multi-Agent RL with Mixed Motives



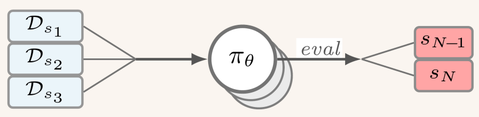
For *each* scenario s_i independently: train only on $\mathcal{D}_{s_i}$, evaluate online on the same scenario. Partner fixed and therefore folded into the environment dynamics.

Evaluation Setting 2: Partner Inference for Offline Multi-Scenario Social RL



Train on the *union* of per-scenario datasets with the scenario label *withheld*; evaluate each scenario separately. A single policy must infer its partner from its interaction history.

Evaluation Setting 3: Zero-Shot Social Generalisation in Offline RL



Train on a *subset* of scenarios; evaluate on a *disjoint* subset. Held-out partners the behaviour policy never interacted with.

Train · Test · Held-out Test

BASELINES

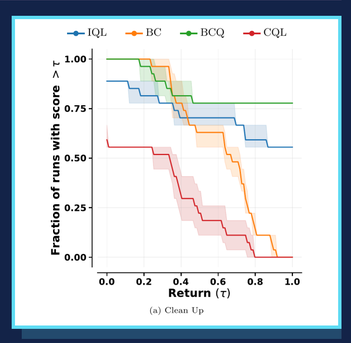
The four canonical offline RL algorithms — the exact tools built for **state–action** shift — each given a **recurrent GRU** over interaction history so partner inference is at least *possible*.

BC · BCQ · IQL · CQL

Behavioural Cloning · **Batch-Constrained Q-Learning** · **Implicit Q-Learning** · **Conservative Q-Learning**

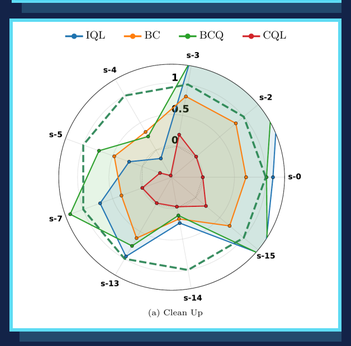
WHAT WE FIND

Setting 1. All four algorithms fall well short of the **p_{90} ceiling**, aggregated across every substrate and scenario — headroom nobody closes, even in this easiest, single-scenario case.



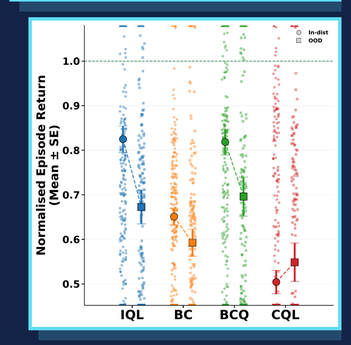
(a) Clean Up

Setting 2. Asymmetric spiders show algorithms failing to adapt to different partners — a **partner-inference failure** the substrate-level aggregate quietly hides.



(a) Clean Up

Setting 3. Every algorithm scores well below its in-distribution level on held-out scenarios — a clear **generalisation gap** that persists no matter which algorithm you pick.



THE TAKEAWAY

Offline RL taught agents to be conservative about **states and actions** they've never seen. It hasn't taught them how to handle **partners they've never met**.

Progress needs better mechanisms for **partner inference** and **social generalisation**. Molten Pot charts a new direction for multi-agent offline RL — with plenty of open problems still to explore.

Supported by **Cooperative AI Foundation** & **AI Safety South Africa**

PAPER · CODE · DATA

Code: MIT · Data: CC BY 4.0