CausalXRL: Explainable Reinforcement Learning through Causal Graph Reasoning

Yanming Zhang Eric Papenhausen Klaus Mueller

Department of Computer Science
Stony Brook University

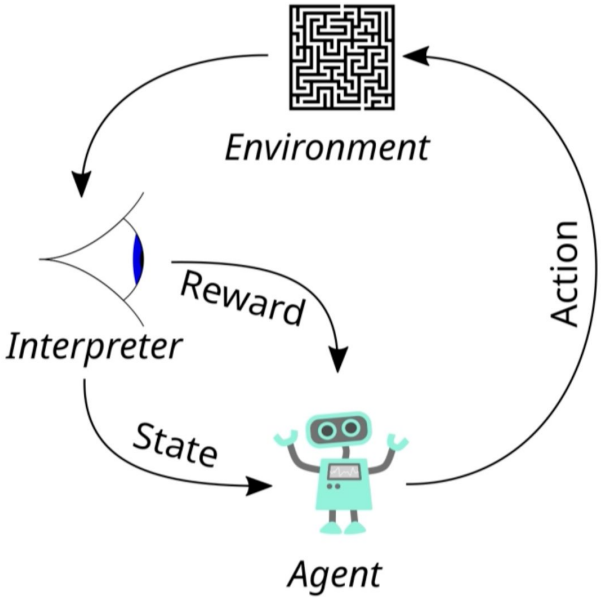
Background: Explainable Reinforcement Learning

AI decision-making is often a black box, especially in high-stakes, complex tasks.

RL: Agents learn through interactions with the environment.

- What has the agent learned?
- Why does it make specific decisions?

XRL: Explains the behavior and decisions of reinforcement learning agents.



RL environment mechanism

Background: RTS games

An example of RTS games

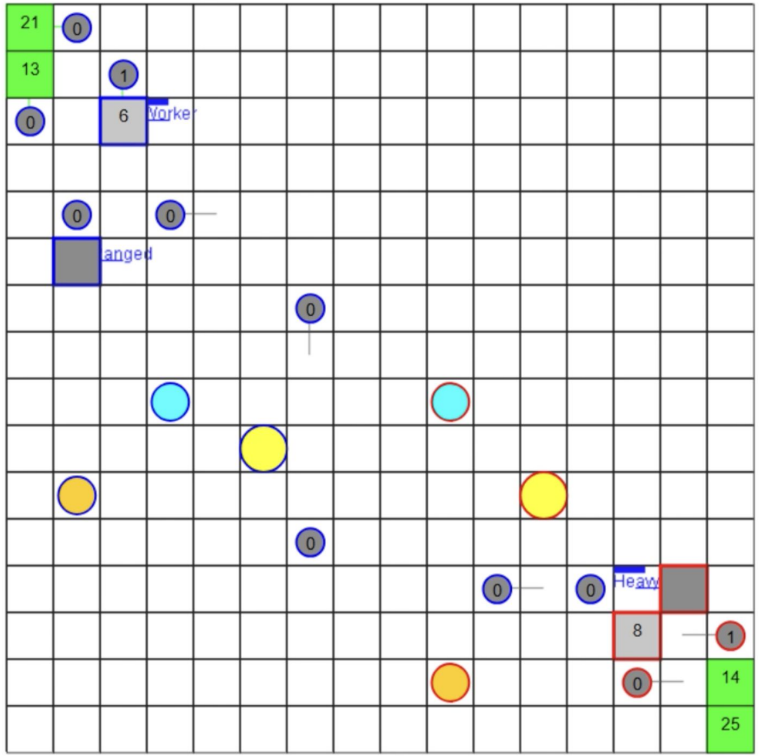
Real-time strategy (**RTS**) games

RTS games involve managing resources, building bases, and commanding armies in real time. They emphasize rapid decision-making and strategic planning rather than turn-based gameplay.

Challenges in Explaining RTS Games

- **High-Dimensional Environments** – Complex state and action spaces.
- **Simultaneous Actions** – Multiple actions occur in real time.
- **Multi-Layered Decisions** – Decisions range from tactical control to strategic planning.

Background: Experiment Environment

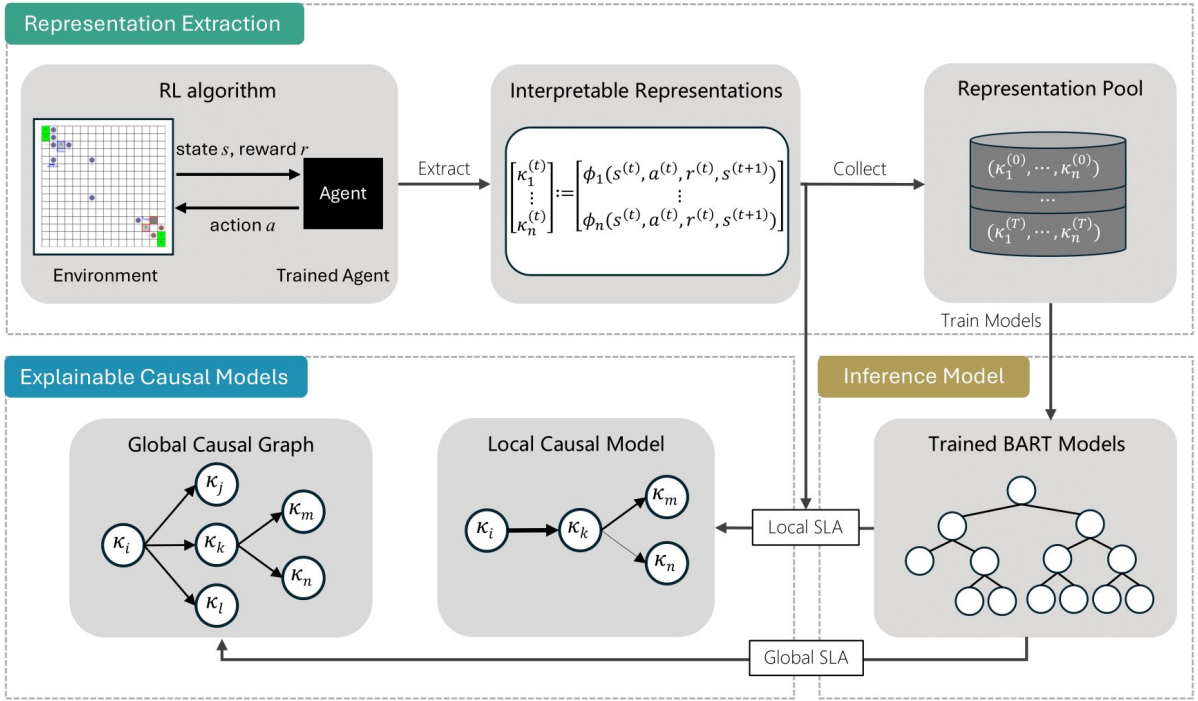


A snapshot of MicroRTS environment

Unit type		Name	Capabilities
Ally units	Enemy units		
25		Resource	A Resource isn't owned by any player, can't perform actions. The number displays the remaining resource.
5	5	Base	A Base has 10 hp, cost 10 resources to build. It can only produce Workers. The number shows resources held by the player(ally/enemy).
		Barrack	A Barrack has 4 hp, costs 5 resources to build. It can produce Light, Heavy, or Ranged units.
		Worker	A Worker has 1 hp, cost 1 resource to build. It can carry 1 resource, move, attack (cause 1 damage), harvest, and return in adjacent cells.
		Ranged	A Ranged has 1 hp, costs 2 resources to build. It can move to adjacent cells or attack locations with $d \leq 3$ (cause 1 damage).
		Light	A Light has 4 hp, costs 2 resources to build. It can only move to or attack (cause 2 damage) adjacent cells.
		Heavy	A Heavy has 4 hp, cost 2 resources to build. It can only move to or attack (cause 4 damage) adjacent cells.

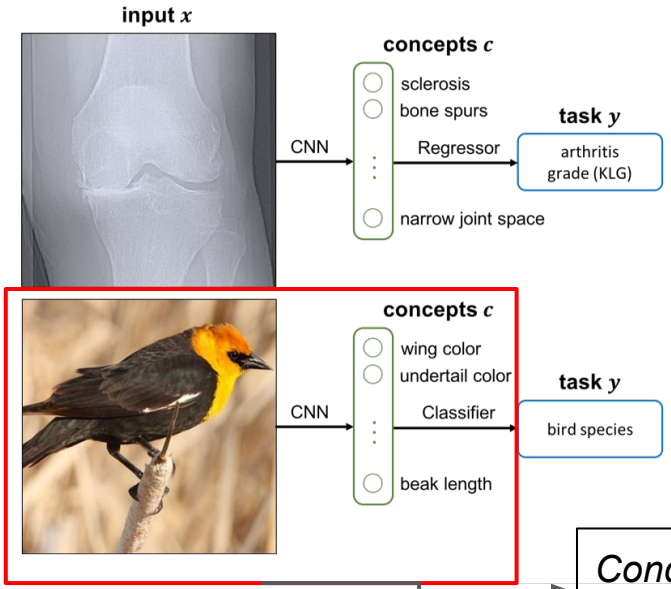
MicroRTS:Type of units and their capabilities

Our Solution: CausalXRL



The architecture of CausalXRL. SLA is the Structural Learning Algorithm.

CausalXRL: Interpretable Feature Extraction



An example of CBM

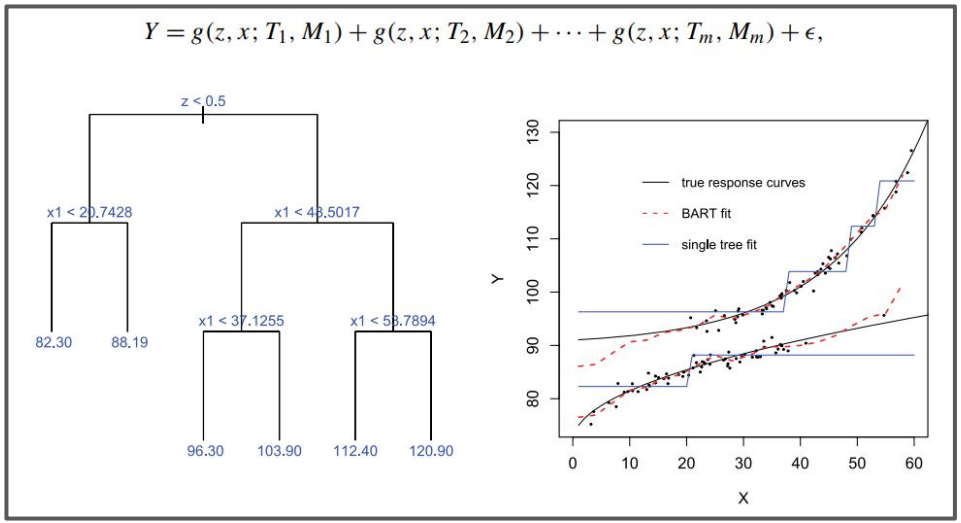
Concept predictor that extracts human understandable concepts

Interpretable Feature Extraction focuses on encoding the state space in a way that is understandable and meaningful to users.

Concept Bottleneck Models (CBMs) are interpretable AI models that first map complex inputs (e.g., images) to human-understandable concepts (e.g., wing color) and then use these concepts to make a final prediction (e.g., bird species).

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020, November). Concept bottleneck models. In International conference on machine learning (pp. 5338-5348). PMLR.

CausalXRL: Train Inference Models



An illustration of BART

Framework Choice

Since we are agnostic to the causal relationships among features, we adopt the Rubin Causal Framework.

Model Choice

We choose Bayesian Additive Regression Trees (BART) as our inference model.

Implementation

- Iterate over all potential treatment–outcome pairs
- Use the remaining features as covariates
- Include the game step as an additional covariate

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics, 20(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

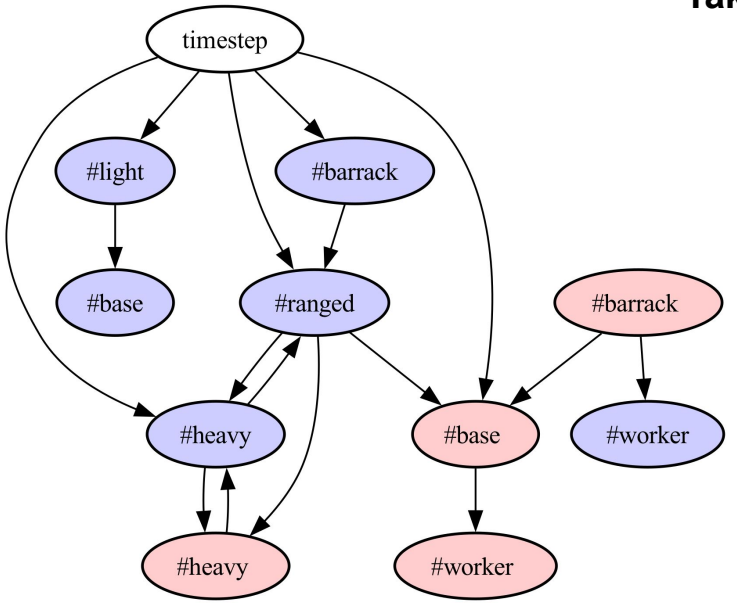
CausalXRL: Infer Global Causal Graphs

Algorithm 1 Global Causal Graph Structure Learning

Input: interpretable representations D.
Output: directed edges $\mathcal{E}$.

/* Preprocessing */
for $t = 0, 1, \ldots, T$ **do**
  $\Delta\kappa^{(t)} = \kappa^{(t+1)} - \kappa^{(t)}$ *// rep. increment*
  $\tilde{\kappa}^{(t)} = \left(z(\kappa^{(t)}_{n\in N}), \kappa^{(t)}_{n\notin N}\right)$ *// standardization*
end for
/* Construct the causal graph */
for $i = 1, 2, \ldots, n$ **do**
  for $j = 1, 2, \ldots, n$ **do**
    $\Delta\kappa_i = f(\tilde{\kappa}_j, \tilde{\kappa}_{-\{i,j\}})$
    $= \sum_k g(\tilde{\kappa}_j, \tilde{\kappa}_{-\{i,j\}}; T_k, M_k) + \varepsilon$
    // fit inference models (BART)
    $\tau^{(t)}_{ji} = f(\tilde{\kappa}_j(1), \tilde{\kappa}_{-\{i,j\}}) - f(\tilde{\kappa}_j(0), \tilde{\kappa}_{-\{i,j\}})$
    // generate CATE samples
    if $P_{2.5}(\tau^{(t)}_{ji}) > 0$ **or** $P_{97.5}(\tau^{(t)}_{ji}) < 0$ **then**
      Append causal link (κ_j, κ_i) to the set $\mathcal{E}$
    end if
  end for
end for

Algorithm to infer global causal model



Inferred global causal model of a gridnet PPO agent

Takeaways

Learn the causal relationships among interpretable features, revealing how an agent's overall behavior influences the environment.

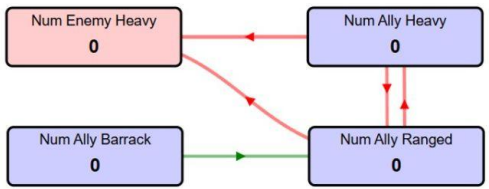
CausalXRL: Infer Local Causal Models

Algorithm 2: Local Causal Model Structure Learning

Input: a single step experience $e^{(t)}$, z-score scalers z, and trained BART models $\mathcal{F}$.
Output: a set of directed edges $\mathcal{E}$ parameterized with weights.

$\kappa^{(t)} = \Phi(e^{(t)})$ // Convert raw exp. into repr.
$\tilde{\kappa}^{(t)} = (z(\kappa^t_{n \in N}), \kappa^t_{n \notin N})$ // standardization
for $i = 1, 2, \ldots, n$ **do**
  for $j = 1, 2, \ldots, n$ **do**
    $\tau^{(t)}_{ji} = f(\tilde{\kappa}_j(1), x_{-\{i,j\}}) - f(\tilde{\kappa}_j(0), x_{-\{i,j\}})$
    // generate CATE samples
    if $P_{2.5}(\tau^{(t)}_{ji}) > 0$ **or** $P_{97.5}(\tau^{(t)}_{ji}) < 0$ **then**
      Append causal link $(\kappa_j, \kappa_i, \bar{\tau}^{(t)}_{ji})$ to the set $\mathcal{E}$
    end if
  end for
end for

Algorithm to infer local causal models

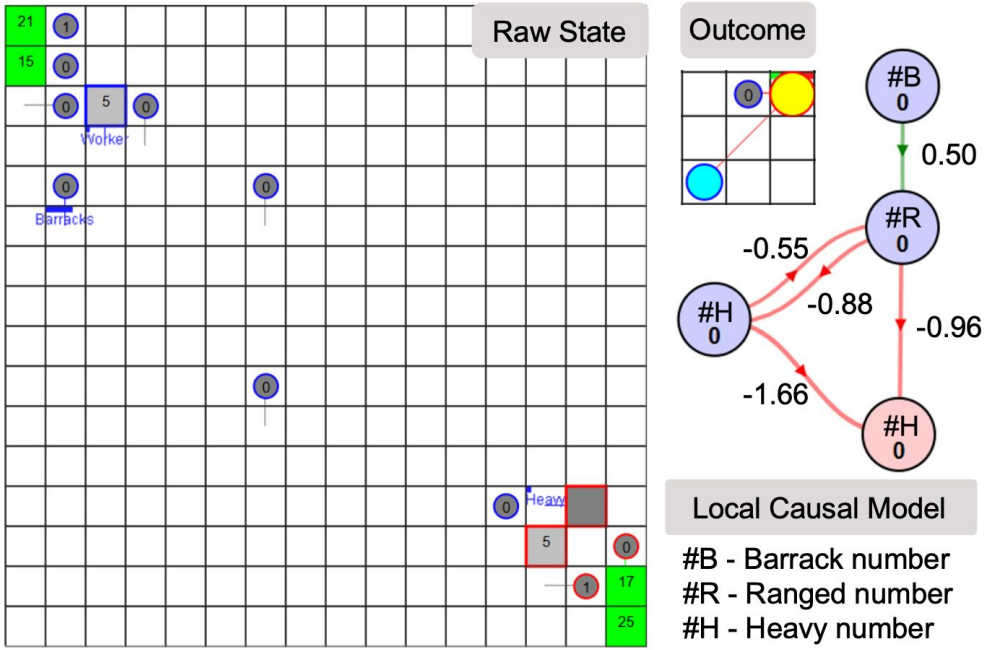


Inferred local causal network at a given state

Takeaways

- Unique causal structures for different strategic states (heterogeneity).
- Macro-level explanations of agent behavior.
- Causal edges capture influence relationships.
- Counterfactual reasoning enables what-if analysis.

Case Study: Local Causal Models



An example of a game state and its corresponding strategic causal network.

- Interpretable Features

  Number of units, (e.g. # of Workers)

- Learned Causal Edges

  Capture how changing one variable affects another while controlling for confounding factors.

- What-if analysis

  Example: If the number of allied Barracks increases, how will the number of enemy Heavies change along the causal chain?

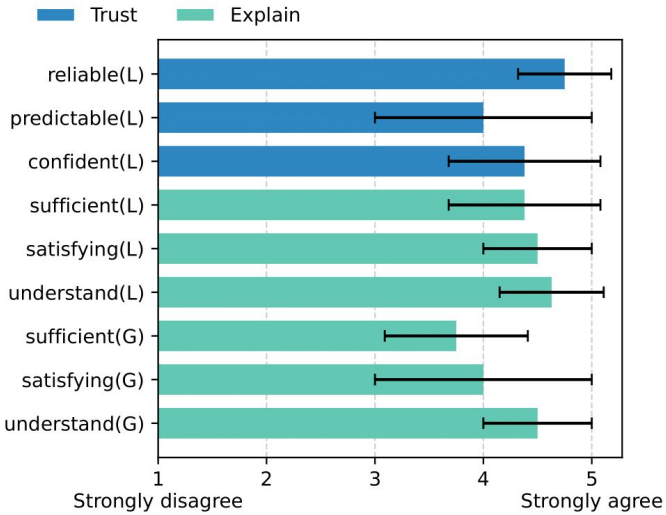
Evaluations

Human Experts Evaluation

Exploratory sessions with participants to study:

1) Whether the proposed causal models improve human understanding of RL agent behavior.

2) Whether improved understanding of agent behavior leads to greater human trust in the agent.



Survey ratings on CausalXRL's explanation quality and participants' trust: Global (G) and Local (L) Models.

Computational Evaluation

Quantitative studies to evaluate fidelity of CausalXRL and Structural Learning Efficiency

RL Env	#Rep	Accuracy (%)			F1-score (%)			Efficiency (s)		
		LM	DT	Our	LM	DT	Our	PC	FCI	Our
CartPole	4	88.5	**94.1**	91.4	90.0	**94.6**	91.6	1.7	0.8	**0.1**
LunarLander	8	81.3	85.1	**88.7**	80.2	84.7	**86.6**	12.6	5.1	**0.3**
PongDuel	12	81.5	88.3	**90.2**	81.1	88.4	**90.7**	23.2	8.7	**0.5**
MicroRTS-Lite	14	75.3	80.3	**86.0**	74.5	82.2	**85.9**	20.9	7.4	**0.5**

Experimental results for behavior prediction accuracy, F1-score, and structural learning efficiency.