Evaluating Saliency Maps Using Interventions

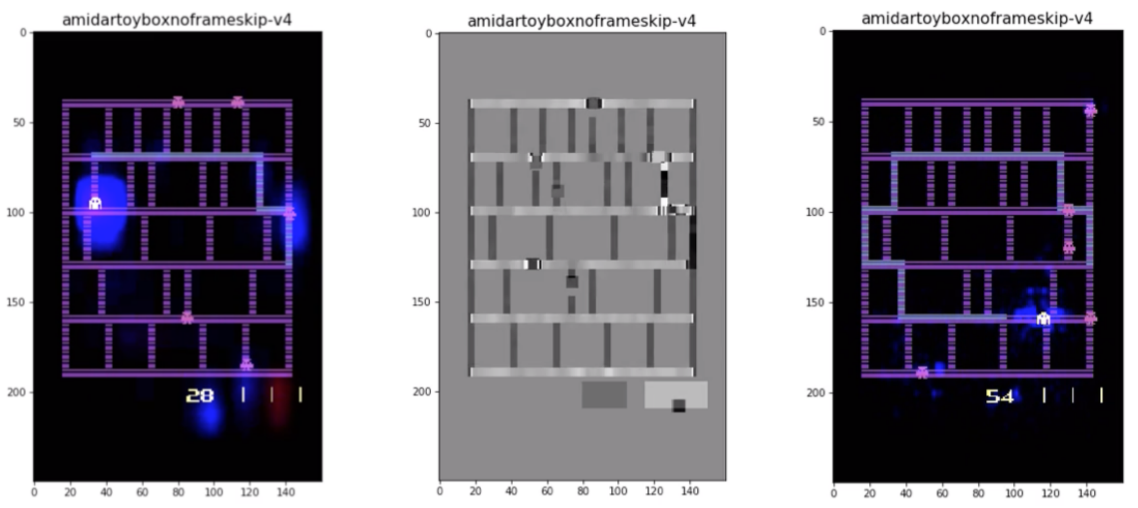
Akanksha Atrey, Kaleigh Clary, David Jensen

University of Massachusetts Amherst

INTRODUCTION

Researchers use saliency maps (SM) to **explain** agent behavior.

Explanations are causal [1].



(a) (b) (c)

Example saliency maps: (a) perturbation; (b) object; and (c) Jacobian.

If saliency maps do not reflect the causal relationships that are assumed by some researchers, **incorrect conclusions** may be drawn from the resulting maps.

CONTRIBUTIONS

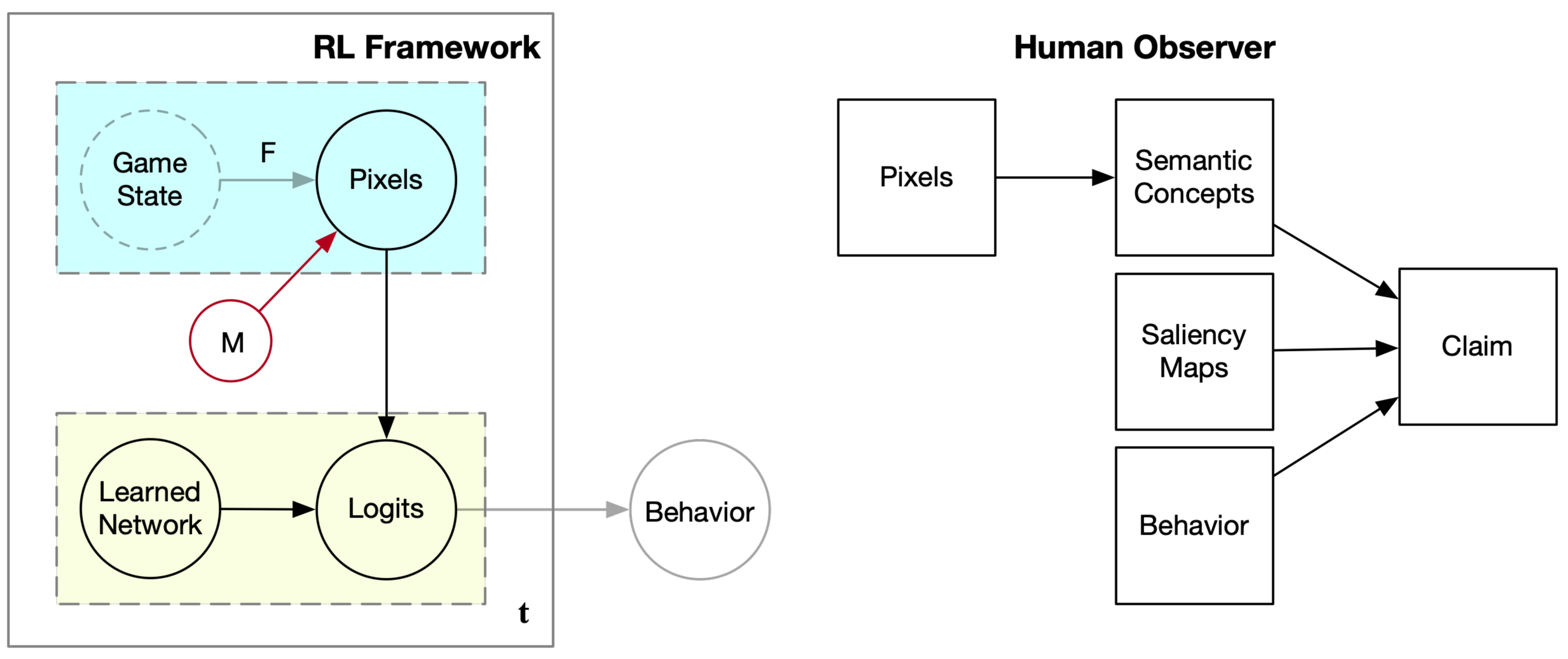
In this work, we develop a methodology grounded in counterfactual reasoning to empirically evaluate the explanations generated using saliency maps in deep RL. Specifically, we:

C1. Survey the ways in which saliency maps have been used as evidence in explanations of deep RL agents.

C2. Describe a new interventional **methodology** to evaluate the inferences made from saliency maps.

C3. Experimentally **evaluate** how well the pixel-level inferences of saliency maps correspond to the semantic-level inferences of humans.

METHODOLOGY



(a) (b)

How should claims be generated?

Concept set {X} is salient → agent has learned representation {R} resulting in behavior {B}.

SURVEY OF USAGE OF SALIENCY MAPS

	Discuss Focus	Generate Explanation	Evaluate Explanation
Jacobian	21	19	0
Perturbation	11	9	1
Object	5	4	2
Attention	9	8	0
Total	46	40	3

Summary of survey of 90 papers with 46 claims drawn from 11 papers that cited and used saliency maps as evidence in their explanations of agent behavior.

Common Pitfalls in Current Usage

Subjectivity. Prior work notes a worrying trend in ML research regarding generating explanations from speculation [6].

Unfalsifiability. Presentation of unfalsifiable interpretations of saliency map patterns.

Assessment of Learned Representations. Limited evidence that (1) salient regions map to learned representations of semantic concepts (e.g., ball, paddle), and (2) the relationships between the salient regions map to high-level behaviors (e.g., channel-building, aiming).

BACKGROUND

Jacobian Saliency. Calculate the gradient of the output with respect to the input [2].

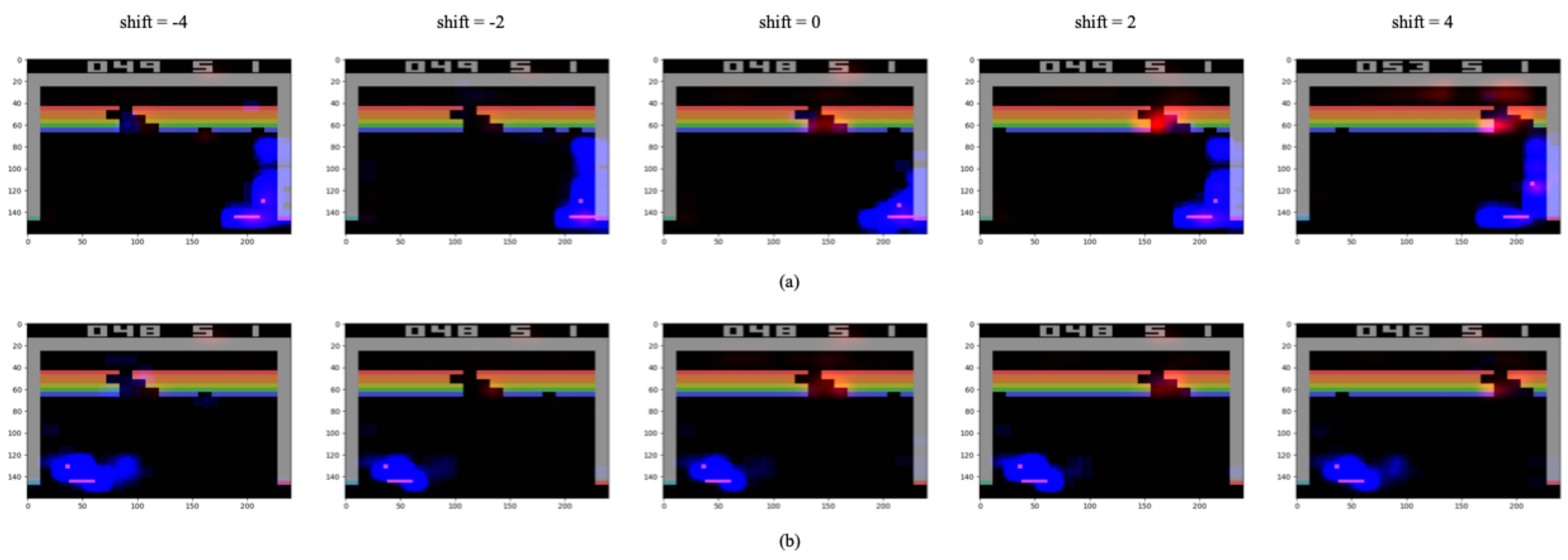
Perturbation Saliency. Perturb the original input image using a Gaussian blur of the image generated by the Hadamard product and measure changes in policy from removing information from a region [3].

Object Saliency. Mask each object with the background color and compute the difference in policy for the unmasked and masked states [4].

Attention Saliency. Use attention activations [5].

EVALUATION OF HYPOTHESES ON AGENT BEHAVIOR

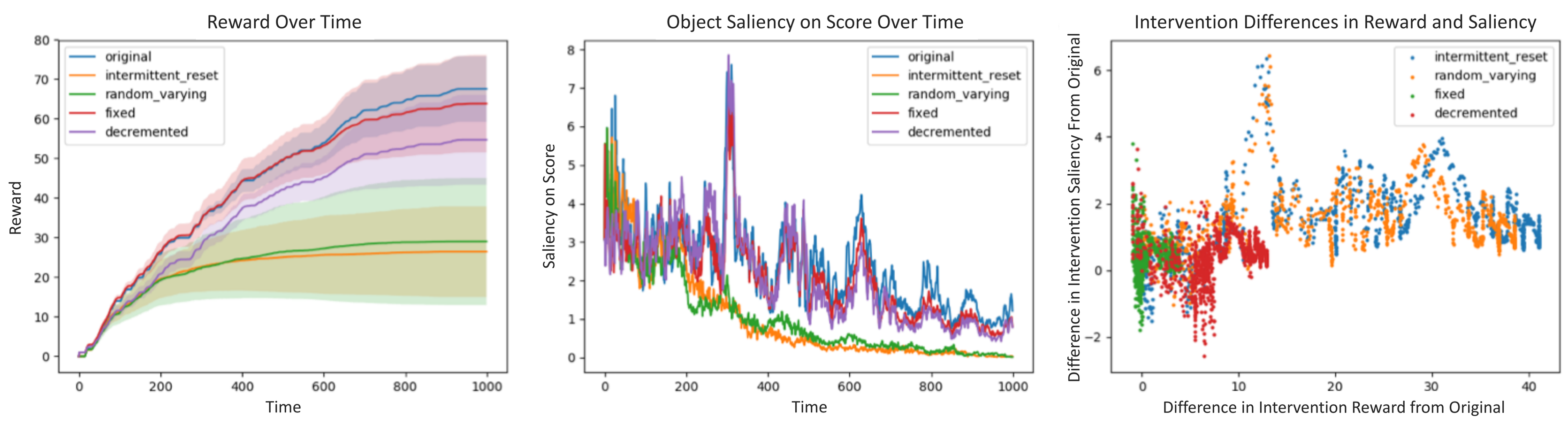
Hypothesis 1: {bricks} are salient ⇒ agent has learned to {identify a partially complete tunnel} resulting in {maneuvering the paddle to hit the ball toward that region}.



Interventions on Breakout: (a) saliency after shifting the brick positions where shift=0 represents the original frame; (b) saliency after shifting the brick positions along with shifting ball and paddle to the left.

Takeaway: The pattern and intensity of saliency around the channel is not symmetric in the reflection interventions.

Hypothesis 2: score is salient ⇒ agent has learned to {use score as a guide to traverse the board} resulting in {successfully following similar paths in games}.

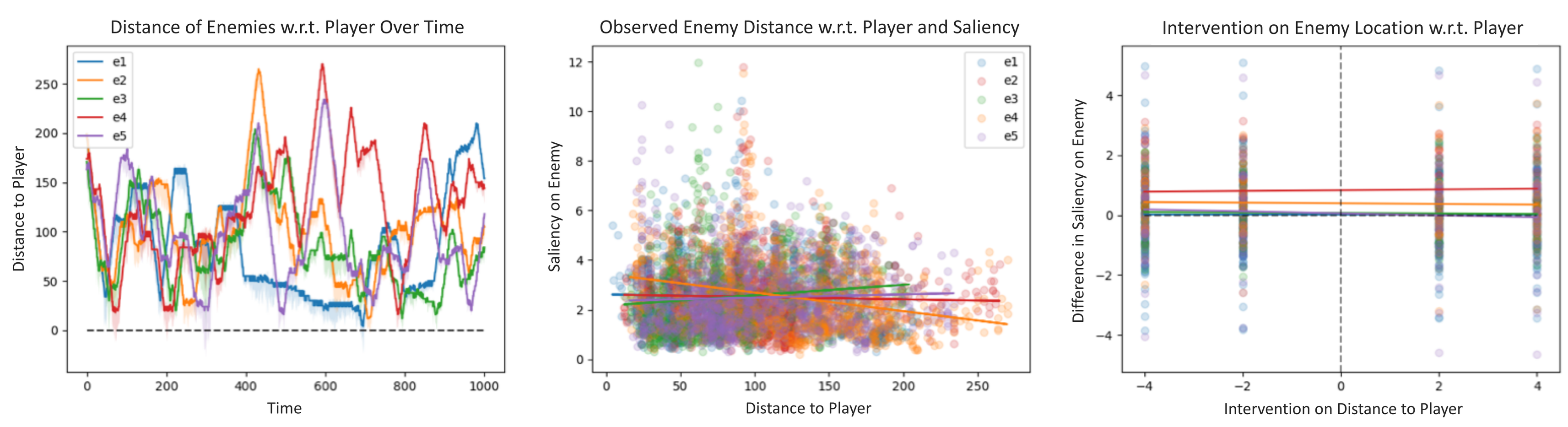


(a) (b) (c)

Interventions on Amidar. (a) reward over time for different interventions on displayed score; (b) saliency on displayed score over time; (c) correlation between the differences in reward and saliency from the original trajectory.

Takeaway: Agent behavior as measured by rewards is underdetermined by salience.

Hypothesis 3: enemy is salient ⇒ agent has learned to {look for enemies close to it} resulting in {successful avoidance of enemy collision}.



(a) (b) (c)

Interventions on Amidar. (a) distance-to-player of each enemy, observed over time, with saliency intensity represented by the shaded region around each line; (b) distance-to-player and saliency, with linear regressions, observed for each enemy; (c) variation in enemy saliency when enemy position is varied.

Takeaway: Spurious correlations can occur between two processes.

REFERENCES

[1] Woodward, James. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[2] Wang, Ziyu, et al. "Dueling network architectures for deep reinforcement learning." *ICML* (2016).

[3] Greydanus, Sam, et al. "Visualizing and understanding atari agents." *ICML* (2017).

[4] Iyer, Rahul, et al. "Transparency and explanation in deep reinforcement learning neural networks." *AAAI* (2018).

[5] Mott, Alex, et al. "Towards Interpretable Reinforcement Learning Using Attention Augmented Agents." *arXiv:1906.02500* (2019).

[6] Lipton, Zachary C., and Jacob Steinhardt. "Troubling trends in machine learning scholarship." *ICML*: The Debates (2018).