Assessing Learned Representations under Open-World Novelty

Kaleigh Clary

University of Massachusetts Amherst
kclary@cs.umass.edu

Abstract

My dissertation research focuses on sequential decision-making (SDM) in complex environments, and how agents can perform well even when novelty is introduced to those environments. The problem of how agents can respond intelligently to novelty has been a long-standing challenge in AI, and poses unique problems across approaches to SDM. This question has been studied in various formulations, including open-world learning and reasoning, transfer learning, concept drift, and statistical relational learning. Classical and modern approaches in agent design offer tradeoffs in human effort for feature encoding, ease of deployment in new domains, and the development of both provably and empirically reliable policies. I propose a formalism for studying open-world novelty in SDM processes with feature-rich observations. I study the conditions under which causal-relational queries can be estimated from non-novel observations, and empirically examine the effects of open-world novelty on agent behavior.

Introduction

AI methods have been successfully applied to construct end-to-end systems for sequential decision-making (SDM) in a variety of domains, including complex human games, healthcare decision support, and advanced real-world tasks. Large-scale models with end-to-end training capabilities have been publicly released or their implementations open-sourced, enabling widespread access and application of generative, discriminative, and decision-making models in the rich-observation setting (e.g., with perceptual features). The generalization, robustness, and behavioral resilience of end-to-end models has been widely studied, and certain machine-learning systems have been identified at risk under both unexpected inputs and adversarial attacks. Given the resource intensity and expense required to produce such agents, if the agent-model supports re-use according to specific performance or behavioral expectations, then we can demonstrate utility toward release of agents to new tasks. Otherwise, it is important to identify the conditions under which models should *not* be used, particularly under safety-critical applications of such technologies (e.g., autonomous vehicles).

In the classical setting, abstractions control the input-space of the agent according to partitions over the inputs. In SDM tasks, environment complexity has previously been coarsely measured using the number of states observed or possible in the environment. Efficiency gains have been measured in abstractions via the total number of states navigable using a reduced model of the world. However, tabular encodings may require significant human effort to produce for real-world systems, and tabular abstractions are rigid and not well-suited to represent most forms of novelty.

My thesis proposes a causal-relational approach to evaluate the bounds of expected behavior of agents developed for sequential decision-making processes, explored through the setting of open-world novelty.

Related Work

In the general case, open-world learning permits any form of novelty. My thesis focuses on forms of novelty which can be expressed as a replacement of any value or set in the relational causal model of the non-novel world. This represents a relational transformation of the space, in line with transformation-based definitions of open-world novelty identified in prior work (Molineaux and Dannenhauer 2022).

The tasks in open-world learning relate to few-shot learning, learning by examples, generalization, transfer learning, covariate shift, domain adaptation, and concept drift. Kirk et al. (2023) introduced the task of compositional generalization, a flexible comparative paradigm for measuring transfer between samples from two different world-systems (e.g., sim-to-real). My thesis offers a specific definition of transformations which relate training and test samples and identify expectations around model re-use.

Boult et al. (2021) introduced a thorough description of the relationship between novelty, the agent, and the environment in the open-world setting. Where the focus in the prior work has been in describing the *effects* of novelty on various aspects of the open-world agent, my thesis seeks to provide a unified formalism for describing the *forms of intervention* expressing and implementing the novelty directly.

Causal models permit certain closure properties over estimation of variables observed in a sample, including in the probabilistic decision-making case (Pearl 2009). Causal-relational methods represent an expressive class of models for causal estimation. However, causal feature selection is known to be a difficult problem in general.

Challenges and Risks in Open-World Novelty

I argue that while interventions may alter or manipulate some variables in the generative distribution of environment transitions, some causal effects may still be estimated from prior sequences of observations in the non-novel world. My thesis offers a formalism for studying the zero-shot transfer of causal queries between non-novel and novel instances. I study forms of transformations induced between sequences of observations under a novelty intervention. My work examines the algebraic relations between these sequences which enable opportunities for causal-relational transfer.

My dissertation research focuses on three primary questions:

RQ1: What are the classes, current practices, and known risks of sequential decision-making in the rich-observation setting under open-world novelty? I propose a taxonomy of novelty interventions organized according to the variables under intervention and their role in the generative distribution of the environment. I propose to conduct a literature review to further develop the taxonomy with respect to prior work. The goal of these efforts is to ensure prior approaches are represented in the taxonomy, and to identify the areas of strong concentration or limited study.

RQ2: Under what conditions can causal-relational models be used to correctly infer causal queries in novel environments? I propose to identify queries which transfer under specific forms of novelty by leveraging algebraic properties of causal and relational models. Causal estimation is enabled under the conditions for identifiability, which can be satisfied by meeting positivity and exchangeability criteria in the sample of observations (Hernán and Robins 2020). I propose a group-theoretic approach to studying relational exchangeability and compositional positivity. The objective is to identify the conditions under which non-novel models can be applied zero- or few-shot to estimate queries under particular forms of novelty intervention.

RQ3: What common training strategies for improving agent representations, skills, or behavior offer advantages in the open-world setting? I propose to empirically study the effects of different methods for training agent-models. I focus on forms of novelty intervention which retain most of the relational structure of the non-novel world. Such designs can empirically measure the effects of, e.g., auxiliary sampling designs during agent-model training, on behavioral outcomes of the agent under open-world novelty.

These research questions center on identifying the opportunities and theoretical requirements for re-use of non-novel model estimates, and their findings inform expectations for agent-model capacity, capabilities in novelty accommodation, efficiency of learning, and safe deployment of AI-supported sequential decision-making systems.

Preliminary Work and Research Timeline

My prior work has focused on measuring, specifying, and evaluating behavior of models (Clary et al. 2018; Tosch et al. 2019) and has shown that certain measures of deep-learning systems (e.g., saliency maps) do not correspond to causal signals of model behavior (Atrey, Clary, and Jensen 2020). My current research studies the reliability and limits of AI systems under unexpected inputs. I have developed a preliminary taxonomy of open-world novelty (**RQ1**) and outlined proof sketches of the algebraic relationships between observed elements in SDM processes (**RQ2**). To empirically evaluate queries (e.g., agent performance) under open-world novelty in the rich-observation setting (**RQ3**), we use environments implemented to enable novelty interventions. I have completed preliminary work using TOYBOX (Foley et al. 2018), an interventional simulator of Atari games. My Timeline is given in Table 1.

Objective	Timeline
Proposal Defense	October 2022
RQ1	November 2022 - January 2023
RQ2	February 2023 - May 2023
RQ3	June 2023 - August 2023
Thesis Writing	September 2023 - November 2023

Table 1: Research Timeline

I have also collaborated in and interfaced with small academic teams to develop and evaluate open-world agents in domains including Cart-Pole, Science Birds, and Monopoly. Each environment supports a variety of interventions.

References

Atrey, A.; Clary, K.; and Jensen, D. 2020. Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning. In *International Conference on Learning Representations*.

Boult, T.; Grabowicz, P.; Prijatelj, D.; Stern, R.; Holder, L.; Alspector, J.; Jafarzadeh, M.; Ahmad, T.; Dhamija, A.; Li, C.; Cruz, S.; Shrivastava, A.; Vondrick, C.; and Scheirer, W. 2021. Towards a Unifying Framework for Formal Theories of Novelty. In *AAAI Conference on Artificial Intelligence*.

Clary, K.; Tosch, E.; Foley, J.; and Jensen, D. 2018. Let's Play Again: Variability of Deep Reinforcement Learning Agents in Atari Environments. In *Critiquing and Correcting Trends in Machine Learning Workshop at NeurIPS 2018.*

Foley, J.; Tosch, E.; Clary, K.; and Jensen, D. 2018. Toybox: Better Atari Environments for Testing Reinforcement Learning Agents. In *Workshop on Systems for ML and Open Source Software at NeurIPS 2018.*

Hernán, M. A.; and Robins, J. M. 2020. *Causal inference: What if.* Boca Raton: Chapman & Hall/CRC.

Kirk, R.; Zhang, A.; Grefenstette, E.; and Rocktäschel, T. 2023. A Survey of Zero-Shot Generalisation in Deep Reinforcement Learning. *Journal of Artificial Intelligence Research*, 76.

Molineaux, M.; and Dannenhauer, D. 2022. An Environment Transformation-based Framework for Comparison of Open-World Learning Agents. *Designing Artificial Intelligence for Open Worlds, AAAI 2022 Spring Symposium.*

Pearl, J. 2009. *Causality: Models, Reasoning and Inference.* New York, NY, USA: Cambridge University Press, 2nd edition.

Tosch, E.; Clary, K.; Foley, J.; and Jensen, D. 2019. Toybox: A Suite of Environments for Experimental Evaluation of Deep Reinforcement Learning. arXiv:1905.02825.