

Explainable Reinforcement Learning Through a Causal Lens

Prashan Madumal , Tim Miller , Liz Sonenberg and Frank Vetere

School of Computing and Information Systems
The University of Melbourne, Australia

pmathugama@student.unimelb.edu.au, {tmiller, l.sonenberg, f.vetere}@unimelb.edu.au

Abstract

Prevalent theories in cognitive science propose that humans understand and represent the knowledge of the world through causal relationships. In making sense of the world, we build *causal models* in our mind to encode cause-effect relations of events and use these to *explain* why new events happen. In this paper, we use causal models to derive causal explanations of behaviour of reinforcement learning agents. We present an approach that learns a *structural causal model* during reinforcement learning and encodes causal relationships between variables of interest. This model is then used to generate explanations of behaviour based on counterfactual analysis of the causal model. We report on a study with **120** participants who observe agents playing a real-time strategy game (Starcraft II) and then receive explanations of the agents' behaviour. We investigated: 1) participants' understanding gained by explanations through task prediction; 2) explanation satisfaction and 3) trust. Our results show that causal model explanations perform better on these measures compared to two other baseline explanation models.

1 Introduction

Driven by lack of trust from users and proposed regulations, there are many calls for Artificial Intelligence (AI) systems to become more transparent, interpretable and explainable. This has renewed the interest in Explainable AI (XAI), explored since the expert systems era [Chandrasekaran *et al.*, 1989]. A key pillar of XAI is *explanation*, a justification given for decisions and actions of the system.

However, much research and practice in XAI pays little attention to *people* as intended users of these systems, [Miller, 2018b]. If we are to build systems that are capable of providing 'good' explanations, it is plausible that explanation models should mimic models of human explanation [De Graaf and Malle, 2017]. Thus, to build XIA models it is essential to begin with a strong understanding on how people define, generate, select and evaluate explanations. This paper contributes to an understanding of the interplay between system provided explanations and human trust in system behaviour.

There is a wealth of pertinent literature in cognitive psychology and social sciences that explores the nature of explanations and how people understand them. As humans, we view the world through a causal lens, where we associate observed events and mechanisms to causal relationships [Slooman, 2005]. We build mental models (causal models) with these relationships to act in the world, to understand new events and also to *explain* events. Importantly, causal models give people the ability to consider *counterfactuals* — events that did not happen, but could have under different situations. Although this notion of causal explanation is also backed by the literature in social psychology [Hilton, 2007], causality is only just starting to become more prevalent in XAI research. Furthermore, compared to the burst of research XAI in planning, machine learning and autonomous agents, explainability in reinforcement learning is hardly explored.

In this paper, we introduce an *action influence* model for reinforcement learning (RL) agents, drawing from insights in cognitive science. We provide a formalization of the model using structural causal models [Halpern and Pearl, 2005]. Our approach differs from previous work in explainability, in that we use causal models to generate *contrastive* explanations for *why* and *why not* questions, which previous models lack. Given assumptions about the direction of causal relationships between variables, we learn the quantitative relationships between variables. We introduce algorithms to generate and select *explanans* (causes that constitutes the explanation of an action) for a given *explanandum* (the action to explain in RL context) from the action influence graph. We define *minimally complete* explanations taking inspiration from social psychology literature [McClure and Hilton, 1997] and through human experiments show that our model performs significantly better than current models in the literature.

We conducted a comprehensive human study using the implemented model for RL agents trained to play the real-time strategy game *Starcraft II*. Experiments were run for **120** participants, in which we evaluate the participants' performance in task prediction [Hoffman *et al.*, 2018, p.12], explanation satisfaction, and trust. Results show that our model performs better than the tested baseline, but its impact on trust is not statistically significant.

The main contributions of this paper are twofold, 1) We introduce and formalise an *action influence* model based on structural causal models and present algorithms to generate

and select explanations; 2) We conduct a between-subject human study to evaluate the proposed model with baselines.

2 Related Work

There exists a substantial body of literature that explores explaining the policies and decisions of Markov Decision Processes (MDP), though most of them do not explicitly focus on reinforcement learning. Elizalde *et al.* [2009] generated explanations by selecting and using ‘relevant’ variables of states of factored MDPs, which were evaluated by relevant domain experts. Taking the long term effect an action has, Khan *et al.* [2009] proposed generating sufficient and minimal explanations for MDPs using domain independent templates.

More recently Wang *et al.* [2016] explored generating explanations in human-robot collaboration scenarios using natural language templates. Policy explanations in human-agent interaction settings has been used to achieve transparency [Hayes and Shah, 2017] and provide summaries of the policies [Amir and Amir, 2018]. Explanation in reinforcement learning has been explored, using interactive RL to generate explanations using instructions of a human [Fukuchi *et al.*, 2017] and to provide contrastive explanations [van der Waa *et al.*, 2018], however, their explanations are not based on an underlying causal model.

Humans understand the world through a causal lens [Slooman, 2005; Pearl and Mackenzie, 2018]. Humans expect XAI systems to have familiar models of explanation [De Graaf and Malle, 2017], thus having causal models of explanation can intuitively provide human-like explanations. Previous work has paid little attention to human-centered explanations and causal explanations in RL agents have not been studied.

Other work on causal explanation has focused on scientific explanations [Salmon, 1984], explanations using causal trees [Nielsen *et al.*, 2012] and causal explanations of fairness [Zhang and Bareinboim, 2018].

Although some recent work has emphasized the importance of causal explanation for explainable AI systems [Miller, 2018b; Klein, 2018; Miller, 2018a], work on generating explanations from causal explanation models for MDPs and RL agents has been absent.

3 Causal Models for Explanations

In this section, we define the *action influence model*, which is based on the notion of *structural causal models* from Halpern and Pearl [2005].

3.1 Preliminaries : Structural Causal Models

Structural causal models (SCMs) represent the world using random variables, divided into exogenous (external) and endogenous (internal), some of which might have causal relationships with each other. These relationships can be described with a set of *structural equations*. Formally, a *signature* \mathcal{S} is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where \mathcal{U} is the set of exogenous variables, \mathcal{V} the set of endogenous variables, and \mathcal{R} is a function that denotes the range of values for every variable $\mathcal{V} \in \mathcal{U} \cup \mathcal{V}$.

Definition 3.1. A *structural causal model* is a tuple $M = (\mathcal{S}, \mathcal{F})$, where \mathcal{F} denotes a set of structural equations,

one for each $X \in \mathcal{V}$, such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$ give the value of X based on other variables in $\mathcal{U} \cup \mathcal{V}$. That is, the equation F_X defines the value of X based on some other variables in the model.

A *context* \vec{u} is a vector of unique values of each exogenous variable $u \in \mathcal{U}$. A *situation* is defined as a model/context pair (M, \vec{u}) . An *instantiation* is defined by assigning variables the values corresponding to those defined by their structural equations. Halpern and Pearl [2005] define a notation of counterfactual models, defined $M_{\vec{X} \leftarrow \vec{x}}$, which means set the values of the vector of endogenous variables \vec{X} to the values \vec{x} and define all successor variables their values based on these new values.

A SCM can be represented by a directed acyclic graph (DAG), in which nodes corresponds to V endogenous variables and edges denote causal relationships.

3.2 Causal Models for Reinforcement Learning Agents

In this section we introduce a general definition of *action influence* models for MDP-based RL agents, which is based on SCMs with addition of actions. A MDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where \mathcal{S} and \mathcal{A} give state and action spaces respectively (here we assume the state and action space is finite and state features are described by a set of variables ϕ); $\mathcal{T} = \{P_{sa}\}$ a set of state transition functions (P_{sa} denotes state transition distribution of taking action a in state s); $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a reward function and $\gamma = [0, 1)$ is a discount factor. Objective of an RL agent is to find a policy π that maps states to actions maximizing the expected discounted sum of rewards. We define a causal explanation model for RL agents as follows.

Formally, a signature \mathcal{S} is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{A})$, in which \mathcal{U}, \mathcal{V} , and \mathcal{R} are as in SCMs, and \mathcal{A} is the set of actions.

Definition 3.2. An *action influence model* for a RL agent is a tuple $(\mathcal{S}, \mathcal{F})$, where \mathcal{S} is as in SCMs, and \mathcal{F} is the set of structural equations, in which we have multiple for each $X \in \mathcal{V}$ — one for each *unique* action set that influences X . A function $F_{X,A}$, for $A \in \mathcal{A}$, defines the causal effect on X from applying action A . The set of *reward variables* $X_r \subseteq \mathcal{V}$ are defined by the set of nodes with an out-degree of 0; that is, the set of sink nodes.

Figure 1 shows a graphical representation of an action influence model as an action influence diagram of an agent for the real-time strategy game *Starcraft II*, with exogenous variables omitted. These models are SCMs except that each edge is associated with an action. Note that for our action influence model, each state variable has a *set* of structural equations: one for each incoming action.

We define the *actual instantiation* of a model M as the model $M_{\vec{V} \leftarrow \vec{S}}$, in which \vec{S} is the vector of state variable values from an MDP.

4 Explanation Generation

In this section we provide definitions and algorithms that generate explanations from our action influence model. At a high level the process of explanation generation has 3 phases:

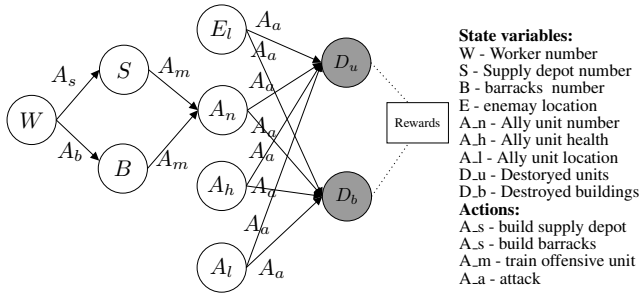


Figure 1: Action influence graph of a Starcraft II agent

Algorithm 1 Causal chain extraction for actions

Input: instantiated causal influence graph M , $explanandum(action) a$

Output: Explanan tuple list $(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_i = \vec{x}_i)$

```

1:  $\vec{X}_r \leftarrow \emptyset$ ; rewards variables
2:  $\vec{X}_h \leftarrow \emptyset$ ; head variables
3:  $\vec{X}_i \leftarrow \emptyset$ ; intermediate variables
4: for  $e \in M.edges$  labelled with action  $a$  do
5:    $\vec{X}_h \leftarrow \vec{X}_h \cup \langle head(e) \rangle$ ; head node of edge  $e$ 
6:    $\vec{X}_r \leftarrow \vec{X}_r \cup \langle sink(e) \rangle$ ; sink/reward nodes
7:    $\vec{X}_p \leftarrow \vec{X}_p \cup \langle getPredecessors(\vec{X}_r) \rangle$ ; predecessors
   nodes of sink nodes
8: end for
9: return list  $(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_i = \vec{x}_i)$ 

```

1) defining the qualitative relationship as an action influence graph; 2) learning the structural equations during RL; 3) then using the SCM to generate *explanans* by instantiating the causal state.

We define an *explanation* as a pair that consist of: 1) an *explanandum*, the event to be explained; and 2) an *explanans*, the subset of causes given as the explanation [Miller, 2018b]. Consider the example ‘Why did you do P ?’ and the explanation ‘Because of Q ’. The *explanandum* is P and *explanans* is Q . Identifying the *explanandum* from a question is not a trivial task. In this paper, we assume that explanandums are of the form ‘Why A ?’ or ‘Why not A ?’, where A is an action. In the context of a RL agent we define a *complete explanan*.

Definition 4.1. An *complete explanan* for an action a under the actual instantiation $M_{\vec{v} \leftarrow \vec{s}}$ is a tuple $(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_i = \vec{x}_i)$, in which \vec{X}_r is the vector of reward variables reached by following the causal chain of the DAG to sink nodes; \vec{X}_h the vector of variables of the head node of action a , \vec{X}_i the vector of intermediate nodes between head and reward nodes, and \vec{x}_r , \vec{x}_h , and \vec{x}_i are the values of these variables under $M_{\vec{v} \leftarrow \vec{s}}$.

Informally, this defines a complete explanan for action a as the complete causal chain from action a to any future reward that it can receive. An algorithm for extracting this is shown in Figure 1.

4.1 ‘Why?’ Questions

Explanatory questions can be broadly divided into three classes: 1) *What*; 2) *How*; and 3) *Why* questions [Miller, 2018b, p. 12]. Lim *et al.* [2009] found that the most demanded explanatory questions are *Why* and *Why not* questions. To this end we focus on explanation generation for *why* and *why not* questions in this paper.

Minimally Complete Explanations

Striking a balance between *complete* and *minimal* explanations depend on the epistemic state of the explainee [Miller, 2018b]. As we are interested in causality, and for simplicity we assume that we know nothing about the epistemic state of the *explainee*. Thus we focus on generating *minimally complete* explanations assuming that the explainee has no prior knowledge about the causal model of the agent.

Recall the Definition 4.1 of *explanans*, a ‘complete’ explanation would include *explanans* of all the intermediate nodes between the head and reward node of the causal chain. Clearly, an explanation with that much *explanans* risks overwhelming the explainee with unnecessary details. For this reason we define *minimally complete* explanations.

McClure and Hilton [1997] show that referring to the goal as being most important for actions. In our causal models, the rewards are the ‘goals’, but these are not meaningful from explanations. As such, we treat the predecessor nodes, which define the immediate causes of the reward, as the ‘end goal’. However, this alone is a longer-term motivation for taking an action. As such, we also include the head node as the immediate reason for doing the action. For this reason we use this model to define our *minimally complete* explanations.

Definition 4.2. A *minimally complete* explanation is a tuple $(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_p = \vec{x}_p)$, in which $\vec{X}_r = \vec{x}_r$ and $\vec{X}_h = \vec{x}_h$ do not change from Definition 4.1, and $\vec{X}_p = \vec{x}_p$ is the vector of variables that are immediate predecessors of any variable in X_r , with \vec{x}_p the values in the actual instantiation.

Informally, a *minimally complete* explanation would have the *explanans* comprising the influenced node of the action (*explanandum*); reward nodes of the causal chain, and the direct causes of the reward nodes.

4.2 ‘Why not?’ Questions

Why not questions let the explainee ask why an event has not occurred, thus allowing *counterfactuals* to be explained. Our model generates counterfactual explanations by comparing causal chains of the actual event occurred and the *explanandum* (counterfactual action). First, we define a counterfactual influence graph that specifies the optimal states under which the counterfactual action B would be chosen, by instantiating all predecessor variables \vec{Z} of the counterfactual action with the current state values (that is, the model $M_{\vec{Z} \leftarrow \vec{s}_Z}$) and then instantiating all successor nodes by simulating using the structural equations. This gives the ‘optimal’ conditions under which we would select counterfactual action B . Then, we find the values in that causal chain that are different to the current state, and compare with the values in the causal chain of *factual* action A .

For readability, in the following definition, we have $\vec{X} = \vec{x}$ to represent the 3-tuple $(\vec{X}_r = \vec{x}_r, \vec{X}_h = \vec{x}_h, \vec{X}_p = \vec{x}_p)$, and similar for $\vec{Y} = \vec{y}$.

Definition 4.3. Given a minimally complete explanation $\vec{X} = \vec{x}$ for action A under the actual instantiation, and a minimally complete explanation $\vec{Y} = \vec{y}$ for action B under the counterfactual instantiation $M_{\vec{Z} \leftarrow \vec{S}_Z}$, we define a *minimally complete contrastive explanan* as the pair $(\vec{X}' = \vec{x}', \vec{X}' = \vec{y}')$ such that \vec{X}' is the maximal set of variables in \vec{X} in which $(\vec{X}' = \vec{x}') \cap (\vec{X}' = \vec{y}') \neq \emptyset$. That is, we only explain things that are different between the actual and counterfactual. This corresponds to the *difference condition* [Miller, 2018a].

We use Algorithm 1 to generate the causal chain, and use the Definition 4.3 to generate counterfactuals. Consider the example below from Figure 1, which is generated from our algorithm (a simple NLP template is used).

Example 4.1. Consider the question asking why a Starcraft II agent built a barracks for training marines, rather than choosing to train marines:

Question Why not do action *build_barracks* (A_b)?
Explanation Because Barrack number (B) is **2** which is optimal: It is more desirable to do action *train_marine* (A_m) to have more Ally units (A_n) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

Here A_b is the *counterfactual case*, and A_m is the actual case with B being the counterfactual values and A_n , D_u and D_b being the counterfactual explanation having the *explanans* of head and reward nodes respectively. There is no extra predecessor node as it is the same as the head node.

4.3 Explanation Selection

Algorithm 1 gives us a set of explanations for a given *explanandum*, and we need to select the most suitable explanation to present to the explainee. We use the following formula to rank the explanations.

Definition 4.4. The ‘best’ explanation from a set of *explanans* is defined as $E = \arg \max \sum |E_o - E_{v_i}|$. That is, the 3-tuple that has the highest combined difference of the current variable values against the optimal variable values of the head, reward and predecessor node is selected.

Our method of selecting the ‘best’ explanation is based on notion of resolving the *cognitive dissonance* of the explainee [Yuan *et al.*, 2011], in that the explanation with the most *surprising* fact will be selected. Most *surprising* explanation would be the one 3-tuple *explanans* that has maximum current variable value difference with their optimal values.

4.4 Learning Structural Causal Equations

Our approach so far relies on knowing the structural model, in particular, to determine the effects of counterfactual actions. *Why not* questions are inherently counterfactual [Balke and Pearl, 1995], and having just the policy of an RL agent is not enough as the counterfactual refers to a *possible* worlds

that did not happen. Consider the Example 4.1, to generate this explanation, optimal/maximum of the state variable (B) is needed in the given time instance.

However, in model-free reinforcement learning, such environment dynamics are not known, and learning a model of the environment is a difficult problem. However, given a graph of causal relations between variables, learning a set of structural equations that are approximate yet ‘good enough’ to give counterfactual explanations may be feasible. To this end, we assume that the a DAG specifying causal direction between variables is given, and learn the structural equations as multivariate linear regression models during the training phase of the agent. In our study in Section 5, as we assume that the relationships between state variables are linear. For other non-linear domains, learning can be done using decisions trees, Bayesian methods, etc.

While this approach may seem similar to learning environment dynamics of model-based RL methods, we only learn the structural equations, not the entire model, and we are only after an approximation that is good enough for explaining instances, thus they can be wrong but still useful for the explainee. Further, specifying the assumptions about the causal direction between variables is a much easier problem to encode by hand, and can be tested with data. In the learning process we take batch samples of 5k steps for each iteration, and feed the state variable values to the set of multivariate regression models obtained from the SCM.

5 Empirical Evaluation: Human Study

A human-grounded evaluation is essential to evaluate explainability of a system, thus we carry out human-subject experiments involving explaining RL agents. We present two main hypotheses for the empirical evaluation as follows; **H1**) Causal model based explanations build better mental models of the agent leading to better **understanding** of its strategies (We make the assumption here that there is no intermediate effect on the mental model from other sources); and **H2**) Better understanding of an agent’s strategies promotes **trust** in the agent.

5.1 Methodology, design and experiments

We chose StarCraft II, a real-time strategy game as the domain of agents. StarCraft II is a popular RL environment [Vinyals *et al.*, 2017] that has a large state space with partial observability. We implemented a Q-learning agent for our experiment that compete with a default agent in game in the default map.

At a high level, our experiment involves, 1) visually displaying how the agent acts in Starcraft II; 2) allowing participants to seek explanations of the agent behaviour; 3) gathering data on participants’ understanding of agent behaviour through prediction tasks; 4) gathering data on explanation satisfaction and participants’ trust of the agent.

To evaluate hypothesis 1 (H1), we use the method of *task prediction* [Hoffman *et al.*, 2018]. Task prediction can provide a quick view of the explainee’s mental model formed through explanations, where the task is for the participant to predict ‘What will the agent do next?’. We use the 5-

point Likert *Explanation Satisfaction Scale* developed recently by Hoffman *et al.* [2018, p.39] to measure the subjective quality of explanations through participant reported values. To evaluate hypothesis 2 (H2), we use the 5-point Likert *Trust Scale* of Hoffman *et al.* [2018, p.49]. These methods will be elaborated in context below.

We recorded a full gameplay video (22 min) with the trained RL agents playing against in-game bot AI. Clips from the gameplay video are the medium of visualizing the agent behaviour to the participant. The experiment has 4 phases.

First phase involves collection of demographic information and training the participants. Using five gameplay video clips, the participant is trained to understand and differentiate actions of the agent. Then, the participant is prompted to identify an action of the agent using another gameplay video.

In the second phase, a clip of the gameplay video (15 sec) is played in a web-based UI, with a textual description of the scene. Participant can select the question type (why/why not) and the action, which together form a question ‘Why/Why not *action A*?’ Then, a textual explanation for the question (using a basic NLP template for the domain) with a figure of the relevant sub-graph of the agent’s causal influence graph is displayed. Explanations are pre-generated from our implemented algorithm and are stored in a web server. The participant can ask **multiple** questions in a single gameplay video. After every gameplay video the participant completes the *Explanation Satisfaction Scale*. This process is repeated so we have data for each participant from five gameplay videos.

The third phase gathers data on the **understanding** that explainee has after seeing the agent play and the explanations. The understanding is measured using the task prediction method as follows: the participant is presented with another gameplay video clip (10 sec), and presented with 3 selections of textual descriptions of what *action* the agent will do in next step; the participant selects an option, and has the option of selecting ‘I dont know’. We expect the participant is projecting forward the *local strategy* of the agent using their mental model formed through explanations (textual explanations of phase 1 are visible to the participant). The participant also provides a free text explanation of why the selection was made and their confidence level. This process is repeated 8 times. In the first 4 task predictions, the agent’s behaviour is similar to the behaviour presented in phase 1, with different variable values. In next 4 tasks, the agent behaviour is novel. In the fourth phase, the participant completes the 5-point *Trust Scale*.

We conducted the experiments on *Amazon MTurk*, a crowd-sourcing platform popular for obtaining high-quality data for human-subject experiments [Buhrmester *et al.*, 2011]. The experiment was fully implemented in an interactive web based environment. Experiment parameters are given below.

The experiment was run with 4 independent variables, one being our causal explanation model and the others being the 3 baselines described below: 1) Gameplay video without any explanation; 2) Relevant variable explanations (explanations are generated using state relevant variables using *template 1* of Khan *et al.* [2009, p.3] and visualized through a state-action graph, e.g ‘Action A is likely to increase *relevant variable*

Model pair	mean-diff	lwr	upr	p-adjusted
R - N	0.433	-1.432	2.298	0.930
D - N	1.666	-0.198	3.532	0.097
C - N	2.400	0.534	4.265	0.006
D - R	1.233	-0.632	3.098	0.316
C - R	1.966	0.101	3.832	0.034
C - D	0.733	-1.132	2.598	0.735

Table 1: Pairwise-comparisons of explanation models of task prediction scores (higher positive diff is better)

P’); 3) Detailed causal influence graph explanations (Causal graph is augmented to include atomic actions, e.g *building_barrack* action is decomposed into *selecting_worker*, *selecting_building_type*, *selecting_build_position*).

We ran experiments for **120** participants, allocated evenly to the independent variables. Each experiment ran approximately 40 minutes. We scored each participant on task prediction, 2 points for a correct prediction; 1 for responding ‘I dont know’ and 0 for an incorrect prediction for a total of 16 points. Scores were tallied. We compensated each participant with 8.5USD, and a bonus of 0.5USD for every points above a score of 13. Of the 120 participants 36 were female, 82 male and 2 were undefined and aged between 19 to 59 ($M=34.2$). Participants had an average self rated gaming experience and Starcraft II experience of 3.38 and 2.02 out of 5 respectively.

5.2 Results

We first present our results on first main hypothesis, corresponding null and alternative hypothesis are, 1) $H_0 : m_C = m_R = m_D = m_N$; 2) $H_1 : m_C \geq m_R$; 3) $H_2 : m_C \geq m_D$; 4) $H_3 : m_C \geq m_N$, where abstract causal explanations (our model), detailed causal explanations, relevant variable explanations, and no explanations are given by C, D, R, and N respectively.

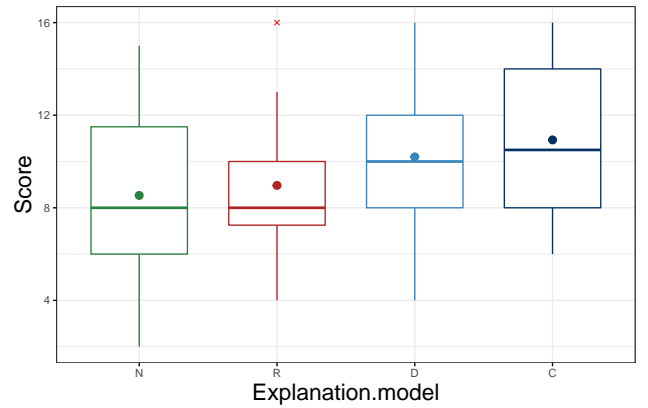


Figure 2: Box plot of task prediction scores of explanation models with means represented as dots (higher is better).

We conduct one-way ANOVA and test for Homogeneity of Variance ($p = 0.65$, Levene’s test) and for normality (Shapiro-Wilk, $p = 0.029 > 0.01$). Figure 2 illustrate the task score variance with explanation models. We obtained a p-value of **0.003** ($M_C = 10.90$, $M_D = 10.20$, $M_R = 8.97$,

Metric	Mdl-pair	Mean-dif	Median-dif	p-val
Complete	C-N	0.707	0.700	0.061
	C-R	0.873	1.000	0.012
Sufficient	C-N	0.746	0.700	0.039
	C-R	1.013	1.000	0.002
Satisfying	C-N	0.633	0.800	0.082
	C-R	0.740	0.700	0.029
Understand	C-N	0.326	0.400	0.497
	C-R	0.400	0.400	0.316

Table 2: Explanation quality survey (likert scale data 1-5)

$M_N = 8.53$), thus we conclude there are significant differences between models on task prediction scores. We performed Tukey multiple pairwise-comparisons to obtain significance between groups. Table 1 shows that the differences between causal explanation model paired with other explanation models is significant for C-R and C-N pairs with p-values of **0.006** and **0.034**. We also calculated the effect of the number of questions might have on the score and found no statistical correlation using a correlation test (number of questions vs score, $p = 0.33$, model C) among same models. We further calculated scores based on 2 = ‘correct’, 0 = ‘incorrect or ‘don’t know’: Results were still significant ($p=0.009$), means (C = 10.9, D = 10.3, R = 8.93, N = 8.73), for model pairs (C-N $p = 0.023$, C-R $p = 0.047$), thus there is no substantial difference in means. Thus we reject H_0 and H_2 and accept all other alternative hypotheses. Results show that causal model explanations lead to significantly better **understanding** of agent’s strategies than 2 baselines we evaluated, specially against previous models of relevant explanations.

We now report our results on the self reported explanation measurements. Figure 3 illustrates the quantitative differences between likert scale and explanation metrics (understand, satisfying, sufficient detail and complete) for aggregated video explanations of explanation models. As before we performed pair-wise ANOVA test, results are summarized in Table 2. Our model obtained statistically significant results and outperformed the benchmark ‘relevant explanation’ (R) for all metrics except ‘Understand’.

We now evaluate the second main hypothesis: Explanation models that achieve better understanding will promote trust in the agent. The obtained p-values for trust metrics *confident*, *predictable*, *reliable* and *safe* were not statistically significant (using pair-wise ANOVA). This indicates that there is *no* significant differences between explanation models to the participant’s perceived level of trust of the agent. Although the difference is not significant we can see causal models have high means and medians (see Figure 4). We conclude that although the explanations and scores are significantly better for our model, to promote trust further interaction is necessary.

We further analysed self reported demographic data to see if there is a correlation between task prediction scores and self reported Starcraft II experience level (5-point Likert). Pearson’s correlation test was not significant ($p = 0.45$) thus we conclude there is no correlation between scores and experience level.

One limitation of our experiment is, while a complex do-

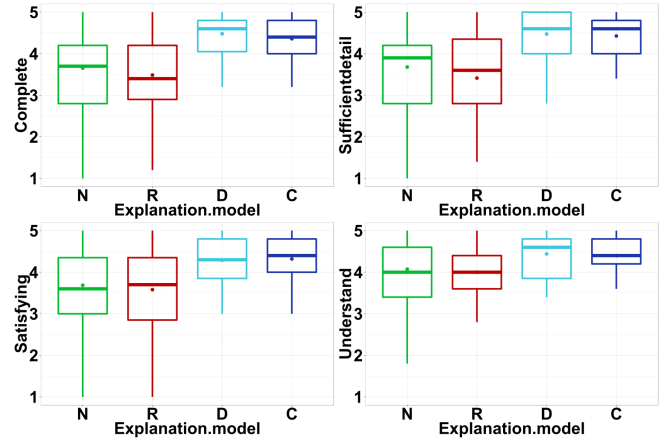


Figure 3: Box plot of explanation quality survey (likert scale 1-5, higher is better, means represented as dots).

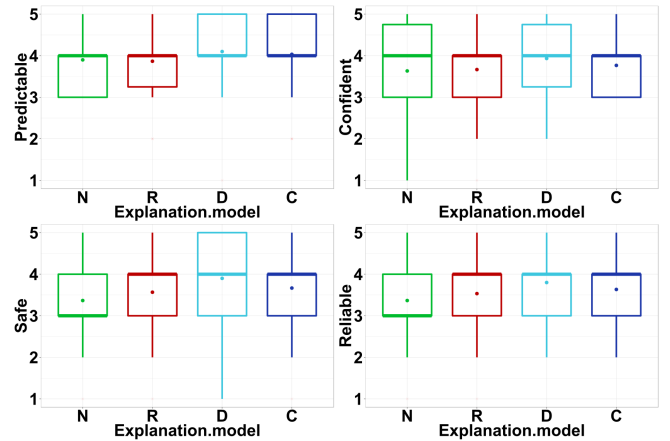


Figure 4: Box plot of trust survey (likert scale 1-5, higher is better).

main comparatively, we made a strong linearity assumption for Starcraft II, which enabled linear regression to learn SCMs for a relatively small number (9) of state variables.

6 Conclusion

In this paper we took inspiration from cognitive science to build an action influence model for reinforcement learning agents. Our approach learns a structural causal (SCM) model during reinforcement learning and has the ability to generate explanations for *why* and *why not* questions by counterfactual analysis of the learned SCM. We conducted a comprehensive human study ($n=120$) to evaluate our model on 1) task prediction, 2) explanation ‘goodness’ and 3) trust. Results show that our model performs significantly better for former 2 evaluation criteria. One weakness of our model is the need of knowing the causal model in advance for the given domain. Future work includes using epistemic knowledge of the explainee to provide explanations that are more targetted, and to developing a suitable abstraction model that can select the relevant level of abstraction for the explainee. We also aim to extend our model to non-factored states in future.

References

- [Amir and Amir, 2018] Dan Amir and Ofra Amir. High-lights: Summarizing agent behavior to people. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2018.
- [Balke and Pearl, 1995] Alexander Balke and Judea Pearl. Counterfactuals and policy analysis in structural models. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 11–18. Morgan Kaufmann Publishers Inc., 1995.
- [Buhrmester *et al.*, 2011] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon’s mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.
- [Chandrasekaran *et al.*, 1989] B Chandrasekaran, Michael C Tanner, and John R Josephson. Explaining control strategies in problem solving. *IEEE Intelligent Systems*, (1):9–15, 1989.
- [De Graaf and Malle, 2017] Maartje M A De Graaf and Bertram F Malle. How People Explain Action (and Autonomous Intelligent Systems Should Too). *AAAI 2017 Fall Symposium on “AI-HRI”*, pages 19–26, 2017.
- [Elizalde and Sucar, 2009] Francisco Elizalde and Luis Enrique Sucar. Expert evaluation of probabilistic explanations. In *ExaCt*, pages 1–12, 2009.
- [Elizalde *et al.*, 2008] Francisco Elizalde, L Enrique Sucar, Manuel Luque, J Diez, and Alberto Reyes. Policy explanation in factored markov decision processes. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM 2008)*, pages 97–104, 2008.
- [Fukuchi *et al.*, 2017] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 97–101. ACM, 2017.
- [Halpern and Pearl, 2005] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 56(4):889–911, 2005.
- [Hayes and Shah, 2017] Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 303–312. ACM, 2017.
- [Hilton, 2007] Denis Hilton. Causal explanation. *Social psychology: Handbook of basic principles*, pages 232–253, 2007.
- [Hoffman *et al.*, 2018] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.
- [Khan *et al.*, 2009] Omar Zia Khan, Pascal Poupart, and James P Black. Minimal sufficient explanations for factored markov decision processes. In *ICAPS*, 2009.
- [Klein, 2018] Gary Klein. Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, 33(2):83–88, 2018.
- [Lim *et al.*, 2009] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128. ACM, 2009.
- [McClure and Hilton, 1997] John McClure and Denis Hilton. For you can’t always get what you want: When preconditions are better explanations than goals. *British Journal of Social Psychology*, 36(2):223–240, 1997.
- [Miller, 2018a] Tim Miller. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163*, 2018.
- [Miller, 2018b] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.
- [Nielsen *et al.*, 2012] Ulf Nielsen, Jean-Philippe Pellet, and André Elisseeff. Explanation trees for causal bayesian networks. *arXiv preprint arXiv:1206.3276*, 2012.
- [Pearl and Mackenzie, 2018] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.
- [Salmon, 1984] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.
- [Sloman, 2005] Steven Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.
- [van der Waa *et al.*, 2018] Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerinx. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706*, 2018.
- [Vinyals *et al.*, 2017] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.
- [Wang *et al.*, 2016] Ning Wang, David V Pynadath, and Susan G Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 109–116. IEEE Press, 2016.
- [Yuan *et al.*, 2011] Changhe Yuan, Heejin Lim, and Tsai-Ching Lu. Most relevant explanation in bayesian networks. *Journal of Artificial Intelligence Research*, 42:309–352, 2011.
- [Zhang and Bareinboim, 2018] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *32nd AAAI Conference on Artificial Intelligence*, 2018.