Distal Explanations for Explainable Reinforcement Learning Agents

Prashan Madumal
University of Melbourne
Victoria, Australia
pmathugama@student.unimelb.edu.au

Tim Miller
University of Melbourne
Victoria, Australia
tmiller@unimelb.edu.au

Liz Sonenberg
University of Melbourne
Victoria, Australia
l.sonenberg@unimelb.edu.au

Frank Vetere
University of Melbourne
Victoria, Australia
f.vetere@unimelb.edu.au

ABSTRACT

Causal explanations present an intuitive way to understand the course of events through causal chains, and are widely accepted in cognitive science as the prominent model humans use for explanation. Importantly, causal models can generate *opportunity chains*, which take the form of 'A enables B and B causes C'. We ground the notion of opportunity chains in human-agent experimental data, where we present participants with explanations from different models and ask them to provide their own explanations for agent behaviour. Results indicate that humans do in-fact use the concept of opportunity chains frequently for describing artificial agent behaviour. Recently, *action influence* models have been proposed to provide causal explanations for model-free reinforcement learning (RL). While these models can generate counterfactuals—things that did not happen but could have under different conditions—they lack the ability to generate explanations of opportunity chains. We introduce a *distal* explanation model that can analyse counterfactuals and opportunity chains using decision trees and causal models. We employ a recurrent neural network to learn opportunity chains and make use of decision trees to improve the accuracy of task prediction and the generated counterfactuals. We computationally evaluate the model in 6 RL benchmarks using different RL algorithms, and show that our model performs better in task prediction. We report on a study with **90** participants who receive explanations of RL agents behaviour in solving three scenarios: 1) Adversarial; 2) Search and rescue; and 3) Human-Agent collaborative scenarios. We investigate the participants' understanding of the agent through task prediction and their subjective satisfaction of the explanations and show that our distal explanation model results in improved outcomes over the three scenarios compared with two baseline explanation models.

KEYWORDS

Explainable AI; Explainable Reinforcement Learning; Human-Agent Interaction; Human-Agent Collaboration

1 INTRODUCTION

Understanding how artificially intelligent systems behave and make decisions has long since been a topic of interest in the research community, and in recent years has resurfaced as 'Explainable AI' (XAI). The ability to provide explanations of the behaviour of these systems is important in critical scenarios where humans need to collaborate with intelligent agents. Often, the success of these collaborative tasks depends on how well the human understands both the long-term goals and immediate actions of the agent.

Explanation models that emulate human models of explanations have the ability to provide intuitive and natural explanations allowing the human a deeper understanding of the agent [1, 8, 25, 32]. As humans, we build *causal* models of the world to encode cause-effect relationships in our mind [29], and use these models to explain why *events* happen. By exploiting causal models, it is possible to provide 'better' explanations to humans. Causal models also enable the generation of *counterfactual* explanations—explanations about events that did not happen but could have under different circumstances [13]. Recently, causal models have been used as *action influence graphs* to generate explanations using *causal chains* for model-free reinforcement learning (RL) agents [22], and have been shown to have subjectively 'better' explanations and better performance in *task prediction* [18] than state-action based explanations [20].

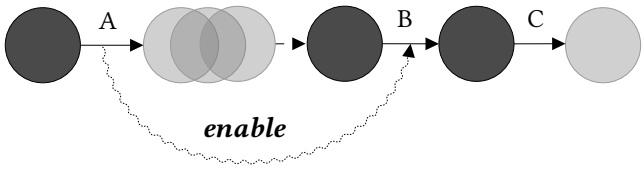

Figure 1: An *opportunity chain* [17], where event *A* *enables* *B* and *B* causes *C*.

Hilton and McClure [16, 17, 23] note that humans make use of *opportunity chains* to describe events through causal explanation. An opportunity chain takes the form of *A* enables *B* and *B* causes *C* (depicted in Figure 1), in which we call *B* the 'distal' event or action. For example, an accident can be caused by slipping on ice which was *enabled* by water from a storm the day before. The notion of opportunity chains is well studied with human experiments in social psychology [16].

Preprint, 2020
2020. https://doi.org/doi

To ground the effect that explanation models have on human explanation, we conduct human-agent experiments where participants received explanations from agents using 3 different explanation models, and were asked to provide their own explanations of the agents' behaviour. Results show that while causality was indeed present, these self-provided explanations predominantly referred to a *future* action that was dependent on the current action, which coincides well with the definition of opportunity chains.

To that end, we propose a *distal* explanation model that can generate opportunity chains as explanations for model-free RL agents. We provide definitions for distal explanations and learn the opportunity chain using a recurrent neural network [28]. A distal explanations by itself would not make a *complete* explanation. For this reason, we use action influence models [22] to get the agent's 'goals'. We further improve upon action influence models by using decision trees to represent the agent's policy.

We computationally evaluate the accuracy of task prediction [18, p.12] and counterfactuals in 6 RL benchmark domains using 6 different RL algorithms, and show that our distal explanation model is robust and accurate across different environments and algorithms. Then we conduct human experiments using RL agents trained to solve 3 different *Starcraft II* [31] scenarios, where agents solve 1) an adversarial task; 2) a search and rescue task; and 3) a human-AI collaborative build task. The human study was run with **90** participants, where we evaluate task prediction [18] and explanation satisfaction. Results indicate that our model performs better than the two tested baselines.

Our main contribution in this paper is twofold: 1) we introduce a distal action explanation model that is grounded on human data; 2) we extend action influence models using decision trees and formalise explanation generation from decision nodes and causal chains. As secondary contributions we also provide the coded corpus of human-agent experiment with **240** explanations and two custom maps that are suited for explainability in the Starcraft II environment.

2 RELATED WORK

Explaining the decisions and policies of autonomous agents has been explored widely, with the focus often given to Markov Decision Process (MDP) based agents. Khan et al. [20] generated *minimal explanations* for MDP agents making use of domain-independent templates and considering the long term impact of an action of the agent. The concept of 'relevant variables' in a factored state of an MDP was exploited by Elizalde et al. [9] to generate explanations, evaluated through domain experts. The effect agent explanations has on 'trust' was examined by Wang et al. [33], while Hayes and Shah [14] sought to improve the transparency by providing policy level explanations for agent based robot controllers.

Explanation in the context of reinforcement learning agents was also explored in recent years. Fukuchi et al. [11] developed explanation generation models using interactive reinforcement learning, aided by human instructions. A decision tree-based approach that is able to generate *contrastive* explanations was proposed by van der Waa et al. [30], though the explanations were not based on a causal model. Providing explanations as policy summarisation has also been explored recently by Amir and Amir [2] and Lage et al. [21].

From early work that represented the agent policy as a decision tree using the 'G' algorithm [6] to more recent work that implemented the idea of representing whole RL agent policies with decision trees for interpretability [27], past literature has explored how decision trees can be used to represent and abstract policies of MDPs. Though to the best of our knowledge previous literature has not studied how decision nodes from a decision tree can be incorporated *with causal chains* to provide explanations.

Some researchers have recently emphasised how humans models of explanations can benefit XAI systems [25] and how humans expect familiar models of explanations from XAI systems [8]. Though some recent progress has been made [22], human-centred computational models of explanation remains largely unexplored.

Hilton and McClure [16, 17, 23] has explored how causal chains of events inform and influence the explanations of humans. *Opportunity chains* can inform the explainee about long term dependencies that events have on each other, where certain events *enable* others. Human experiments have also been carried out that investigate he effects of opportunity chains on explanation [17].

Our proposed *distal explanation* model take insights from social psychology literature to combine *opportunity chains* with *causal* explanations. To the best of our knowledge this is the first of such model in the context of explainable reinforcement learning agents.

3 INSIGHTS FROM HUMAN EXPLANATIONS

In this section, we discuss insights we can gain from human models of explanation in literature. We then ground these models in data by conducting a human-agent experiment.

3.1 Human Models of Explanation

Causality is a recurring concept in explanation models of social psychology and cognitive science literature [15, 19]. Using causal models as the basis for explanation seems natural and intuitive to humans [29], since we build causal models to represent the world and to reason about it. Thus, it is plausible that, when used in intelligent agents, causal models have the ability to provide 'good' explanations to humans.

Importantly, causal models consist of *causal chains*. A causal chain is a path that connects a set of events, where a path from event A to event B indicates that A has to occur before B [25] (we use event and action interchangeably in the paper). Hilton and McClure [17] define 5 types of causal chains that lead to 5 different types of explanations. Hilton et al. [16] have explored how humans select different causal chains to provide explanations through human experiments. We conduct a similar study to gain insights from human models of explanation in a human-agent setting, and report results below.

3.2 Grounding in Human-Agent Experimental Data

We conducted a human-agent study with 30 participants. Participants were shown reinforcement learning agents playing the game Starcraft II, and were given explanations of agents' actions using one of 3 different explanations models: 1) No explanations, just visual description of the agent's behaviour; 2) State-action based explanations [20]; and 3) Causal explanations [22]. Participants

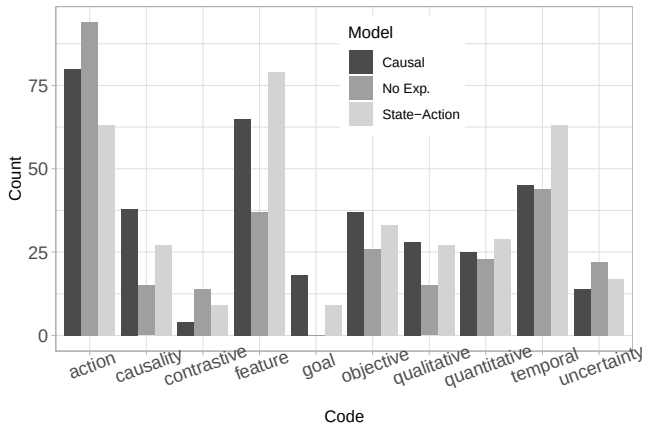

Figure 2: Codes and their frequencies of 240 human explanations of reinforcement learning agents (that were using 3 different explanation models)

were then shown new agent behaviour and were asked to formulate their own explanations about the agent, repeated for 8 rounds. We obtained a total of **240** explanations.

We use grounded theory [7] to code the data and to identify concepts. Figure 2 shows the frequencies of 9 codes across the 3 explanation models of the RL agents. Participants referred to 'actions' and 'features' of the agent the most, and often included the 'objective' or the 'goal' of the agent in the explanation, which is present in action influence models [22]. Most importantly, the third most frequent code is 'temporal', in which participants refer to future actions the agent will take in the explanation (i.e. distal actions). For example, consider an explanation from the data corpus, ''The AI will want to have barracks so that it can then train soldiers to engage in attacks. It will want to progress". Here, the participant's explanation contains the distal action 'train soldiers' which is *enabled* by 'have barracks'. 'Causality' is also present in the explanations, interestingly even in 'No explanation' and State-action based explanation models. This suggests that humans frequently associate causal relationships when generating explanations. Our human-agent experimental data reaffirm the presence of opportunity chains in causal chains [16], and show that these are frequently used to express how future actions are dependent on current actions of agents. This motivates our proposal of a new distal explanation model.

4 DISTAL EXPLANATION MODEL

Before formalising the distal explanation model, we introduce explanations from *action influence* models for MDP based reinforcement learning agents. In the following sections we use the adversarial scenario (discussed at length in Section 5.1) of the Starcraft II environment as a running example to unravel definitions. Figure 3 b) illustrate the action influence graph of this scenario.

4.1 Preliminaries

An 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); $\mathcal{T} = \{P_{sa}\}$ gives a set of state transition functions where 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 $\gamma = [0, 1)$ gives a discount factor. The objective of a reinforcement learning agent is to find a policy π that maps states to actions maximizing the expected discounted sum of rewards.

The *actual instantiation* [22] of a model M is defined as $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$, in which $\vec{S}$ is the vector of state variable values from an MDP and $\mathcal{V}$ gives the set of endogenous variables of the action influence model. A *counterfactual instantiation* [22] for a counterfactual action B is a model $M_{\vec{Z} \leftarrow \vec{S_Z}}$, where $\vec{Z}$ gives the instantiation of all predecessor variables of action B with current state values *and* the instantiation of all successor nodes (of B) of the causal chain by forward simulation.

Using above definitions of instantiation, minimally complete explanations for 'why' and 'why not' questions for action influence models are defined as below.

Definition 4.1. A *minimally complete* explanation [22] for a *why* question is a tuple $\left(\vec{X_r} = \vec{x_r}, \vec{X_h} = \vec{x_h}, \vec{X_p} = \vec{x_p}\right)$, in which $\vec{X_r}$ is the vector of reward variables reached by following the causal chain of the graph to sink nodes; $\vec{X_h}$ the vector of variables of the head node of action a, and $\vec{X_p} = \vec{x_p}$ is the vector of variables that are immediate predecessors of any variable in X_r within the causal chain, with $\vec{x_p}$ the values in the actual instantiation.

In the following definition, $\vec{X} = \vec{x}$ is used to represent the tuple $\left(\vec{X_p} = \vec{x_p}, \vec{X_h} = \vec{x_h}, \vec{X_r} = \vec{x_r}\right)$, and similar for $\vec{Y} = \vec{y}$ for readability.

Definition 4.2. 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 explanation* [22] for a *why not* question is a tuple $(\vec{X'} = \vec{x'}, \vec{Y'} = \vec{y'}, \vec{X_r} = \vec{x_r})$ such that $\vec{X'}$ is the maximal set of variables in $\vec{X}$ in which $(\vec{X'} = \vec{x'}) \cap (\vec{Y'} = \vec{y'}) \neq \emptyset$, where $\vec{x'}$ is then contrasted with $\vec{y'}$. That is, we only explain things that are different between the actual and counterfactual. This corresponds to the *difference condition* [24]. And $\vec{X_r}$ gives the reward nodes of action A.

Informally, a contrastive explanation is generated by extracting the actual causal chain for the taken action A, and the causal chain for the counterfactual action B, and then finding the differences between those causal chains. For example, the explanation for the question 'Why not build barracks (A_b from Figure 3 b))' can be generated by: 1) getting the causal chain of the actual action A_s (highlighted in black in Figure 3 b)); 2) getting the causal chain for the counterfactual action A_b, which gives the nodes $B \rightarrow A_n \rightarrow [D_u, D_b]$; and 3) contrast values of S with the actual and counterfactual instantiation.

While being an important step towards causal explanation of RL agents, action influence models lack the mechanisms to generate distal explanations in their current form. We refer the reader to [22] for a more complete discussion of action influence models.

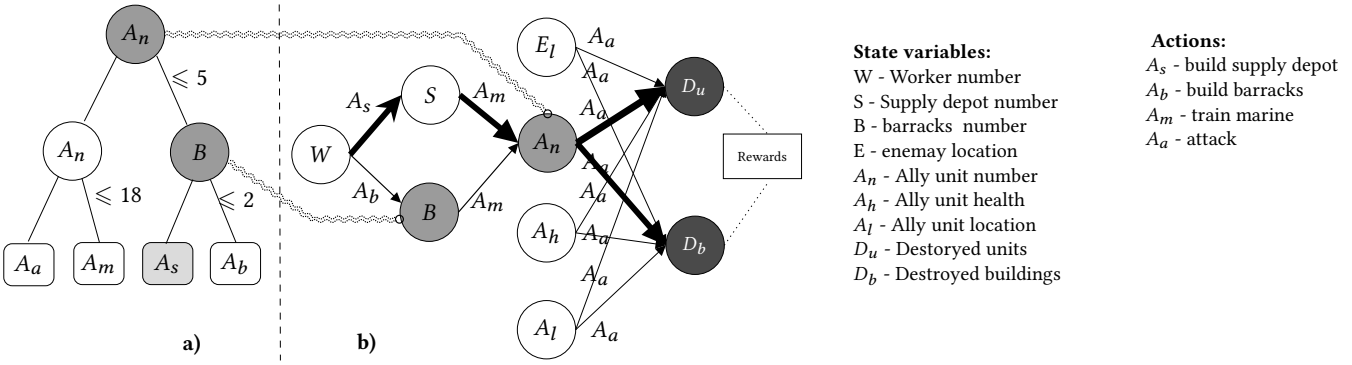

Figure 3: Generating explanations by mapping decision nodes to causal chains.

4.2 Causal Explanations from Decision Trees

Although causal explanations from action influence models have been shown to perform better than state-action based [20] explanation models, the use of *structural equations* yield poor results [22] in computational task prediction accuracy. In place of structural equations that approximate the causal effect between two feature variables, we propose to use decision nodes from a decision tree to generate explanations from causal chains.

The *distal explanation* model we introduce use decision nodes of a decision tree that represent the agents' complete policy to generate explanations with the aid of causal chains from an action influence model. Let $\widehat{\mathbb{T}}$ be a decision tree model. In each episode at the training of the RL agent, we perform experience replay [26] by saving $e_t = (s_t, a_t)$ at each time step t in a data set $D_t = \{e_1, ..., e_t\}$. Drawing uniformly from D as mini-batches, we train $\widehat{\mathbb{T}}$ using input $x = \vec{s}$ and output $y = \vec{a}$. Clearly, explanations generated from a unconstrained decision tree can overwhelm the explainee, as these produce a large number of decision nodes for a question. Thus we limit the growth of $\widehat{\mathbb{T}}$ by setting the max number of leaves to the number of actions in the domain (i.e. the leaves of the trained $\widehat{\mathbb{T}}$ will be the set of actions of the agent). We later show in evaluation, that this hardly affects the task prediction accuracy compared to a depth unconstrained decision tree. To get the decision nodes of $\widehat{\mathbb{T}}$ in state S_t, we simply traverse the tree from the root node until we reach a leaf node and get the nodes of the path. The decision tree of the Starcraft II adversarial task is given in Figure 3 a), with the decision nodes A_n and B for the action A_s. Each decision node maps to a feature variable of the agent's state. Figure 3 shows how the decision nodes are mapped to the action influence graph, in the Starcraft II adversarial scenario.

In the context of an RL agent, we introduce a new definition of *minimally complete* explanations using decision nodes for 'why' questions below.

Definition 4.3. Given the set of decision nodes $\vec{X_d} = \vec{x_d}$ for the action a from a decision tree $\widehat{\mathbb{T}}$, we define a *minimally complete* explanation for a *why* question as a pair $\left(\vec{X_r} = \vec{x_r}, \vec{X_n} = \vec{x_n}\right)$, in which $\vec{X_r}$ is the vector of reward variables reached by following the causal chain of the graph to sink nodes; $\vec{X_n}$ is such that $\vec{X_n}$ is the maximal set of variables in which $\vec{X_n} = (\vec{X_a} = \vec{x_a}) \cap (\vec{X_d} = \vec{x_d})$, where $\vec{X_a}$ is the set of intermediate nodes of the causal chain of action a, with $\vec{x_r}$, $\vec{x_a}$ and $\vec{x_d}$ giving the values under the actual instantiation $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$.

In the Starcraft II scenario, for the question 'why action A_s', we generate the minimally complete explanation by first finding the decision nodes for action A_s, shown as medium grey nodes in Figure 3 a). Then finding the causal chain of action A_s (given by the bold path in Figure 3). And finally getting the common set of nodes from the causal chain and the decision nodes (B in Figure 3) and appending the reward nodes (D_u and D_b).

4.3 Contrastive Explanations from Counterfactuals

Counterfactuals explain events that did not happen—but could have under different circumstances. Counterfactuals are used to describe events from a 'possible world' and to contrast them with what happened in actuality. Embedding these counterfactuals in explanations can make the explanation more meaningful [5]. Naturally, an explanation given to a 'why not' question should compare the counterfactuals with the actual facts to form a *contrastive explanation* [24, 25]. For this reason, we concern ourselves with generating contrastive explanations from decision nodes and causal models.

We generate the counterfactual decision nodes using Algorithm 1, in which we find the decision nodes of the counterfactual action b by changing the decision boundary of the actual action b in the decision tree. We can now define *minimally complete contrastive* explanations for 'why not' questions using these counterfactual decision nodes.

Definition 4.4. Given the set of decision nodes $\vec{X_d} = \vec{x_d}$ for the action a from a decision tree $\widehat{\mathbb{T}}$, a *minimally complete contrastive* explanation for a *why not* question is a pair $\left(\vec{X_r} = \vec{x_r}, \vec{X_{con}} = \vec{x_{con}}\right)$, in which $\vec{X_r}$ is same as in Definition 4.1; $\vec{X_{con}}$ is such that $\vec{X_{con}}$ is the maximal set of variables in which $\vec{X_{con}} = (\vec{X_b} = \vec{x_b}) \cap (\vec{X_c} = \vec{x_c})$, where $\vec{X_b}$ gives the set of intermediate nodes of the causal chain of the counterfactual action b, and $\vec{X_c}$ is generated using the Algorithm 1. Values $\vec{x_r}$, $\vec{x_c}$ are contrasted using the actual instantiation $M_{\vec{\mathcal{V}} \leftarrow \vec{S}}$ and counterfactual instantiation $M_{\vec{Z} \leftarrow \vec{S_Z}}$.

Algorithm 1 Generating Counterfactuals

Input: causal model $\mathcal{M}$, current state S_t, trained decision tree $\widehat{\mathbb{T}}$, *actual* action a

Output: contrastive explanation t $\vec{X_c}$

$\vec{X_d} \leftarrow \widehat{\mathbb{T}} \cdot traversetree(a)$;vector of decision nodes of a from $\widehat{\mathbb{T}}$
$\vec{X_c} \leftarrow []$;vector of counterfactual decision nodes.
for every $D \in \vec{X_d}$ **do**
  $x_d \leftarrow D \cdot decisionNodeValue()$; decision boundary value of D
  $x_m \leftarrow D \cdot moveBoundary(x_d)$; boundary value changed by a Δ.
  $S_t m \leftarrow S_t \cup x_m$; modify the corresponding state feature variables with the new x_m.
  $\vec{X_c} \leftarrow \vec{X_c} \cup \widehat{\mathbb{T}} \cdot predict(S_t m)$; get the counterfactual decision nodes by getting the counterfactual action and then traversing the tree.
end for
return $\vec{X_c}$

As before, we explain Definition 4.4 using the adversarial Starcraft II task. Consider the question 'Why not action A_b', when the actual action is A_s, for which the explanation is generated as follows. We first get the decision nodes A_n and B having <= 5 and > 2 as the decision boundaries respectively. Then each decision boundary value starting with the node closest to the leaf node, is moved by a small Δ amount 0.01 and applied as the new feature value in the current state of the agent (B feature value will change to 1.99). We use this new state to predict the counterfactual action as A_b from the decision tree, and to get the counterfactual decision nodes (which remains the same). Next, we get the intersection of nodes in the causal chain of the counterfactual action A_b ($B \rightarrow A_n \rightarrow [D_u, D_b]$) with $\vec{X_c}$, which gives B as $\vec{X_{con}}$ with the actual value 3 and counterfactual value 1.99. Finally, these values are contrasted and appended with the reward nodes of the causal chain of A_b to generate the explanation.

4.4 Learning Opportunity Chains

In the context of reinforcement learning, we define a 'distal action' as the action that depends the most on the execution of the *current* action of the agent. The agent might not be able to execute the distal action unless some other action was executed first (i.e. *some* actions 'enable' the execution of other actions). For example, in the Starcraft II domain, the action 'train marines' cannot be executed until 'build barracks' action is executed. While it is possible to extract distal actions from environment dynamics and pre-conditions in a model-based system, for model-free RL agents, this remains a challenge. However, for the purpose of explanation, it is possible to provide an approximation and predict the distal action.

We use a many-to-one recurrent neural network (RNN) [28] as our prediction model $\widehat{\mathbb{L}}$ to approximate the distal action given a sequence of previous states and actions of the agent. We implement $\widehat{\mathbb{L}}$ with a fully connected hidden layer of 10 units, and a batch size of 100. For training data, we use the dataset D_t discussed in section 4.2. We define a sequence as a state-action trace that ends in an action that is one of the last actions in a causal chain (e.g. in Figure 2, the last action of all causal chains in the 'attack' action). The output of the model $\widehat{\mathbb{L}}$ will be the distal action and its expected cumulative reward. Note that even though we used an RNN to implement the prediction model, it is entirely possible to use other models to approximate the distal action. With the distal action prediction model $\widehat{\mathbb{L}}$ in hand, we now define *minimally complete distal* explanations for 'why' and 'why not' questions that incorporate causal nature to the explanations.

Definition 4.5. Given a *minimally complete contrastive* explanation, current action a and a prediction model $\widehat{\mathbb{L}}$, a *minimally complete distal* explanation is a tuple $\left(\vec{X_r} = \vec{x_r}, \vec{X_{con}} = \vec{x_{con}}, a_d\right)$, in which $\vec{X_r}$ and $\vec{X_{con}}$ do not change from Definition 4.2; and a_d gives the distal action predicted through $\widehat{\mathbb{L}}$ such that $a_d \in A \cap A_c$, where A is the action set of the agent and A_c gives the action set of the causal chain of current action a.

Informally, this simply prepends the predicted distal action to a minimally complete contrastive explanation generated through Definition 4.4 if the distal action exists in the causal chain of the current action. Consider the example 'Why not action A_b, when the actual action is A_m. This would yield the counterfactual decision node A_n with the actual value 10 and the counterfactual value 5. When the predicted distal action is A_a, we can generate the below explanation text using a simple natural language template. The causal explanation is generated with Definition 4.2 while the distal explanation is generated through Definition 4.5.

Causal Explanation: Because it is more desirable to do the 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).

Distal Explanation: Because ally unit number (A_n) is less than the optimal number 18, it is more desirable do the action train marine (A_m) to *enable the action* attack (A_a) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

Note that the Definition 4.5 can also be used in conjunction with the Definition 4.3 to generate distal explanations for 'why' questions.

4.5 Computational Evaluation

We use 5 OpenAI benchmarks [3] and the adversarial Starcraft II scenario (discussed in Section 5.1) to evaluate the task prediction [18] accuracy of our distal explanation model and compare against action influence models [22] as a baseline. Task prediction can be used to predict what the agent will do in the next instance, and is a useful measure that can give insights on how well the underlying model represent the problem.

We choose the benchmarks to have a mix of complexity levels and causal graph sizes (given by the number of actions and state variables). We train the RL agents using different types of model-free RL algorithms (see Table 1), using a high performance computer cluster node with 2 Nvidia V100 GPUs, 56GB of memory and 20 core CPU with 2.2GHz speed. All agents were trained until the reward threshold (to consider as 'solved') of the environment specification is reached.

We evaluate two versions of the distal explanation model, where one is a based on a depth limited decision tree with the number of

Env - RL		SE - Accuracy (%)			DP - Accuracy (%)	
	Size	LR	DT	MLP	*DP*	DP_n
Cartpole-PG	4/2	83.8	81.6	86.0	**96.83**	97.10
MountainCar-DQN	3/3	69.7	57.8	69.6	**88.66**	86.75
Taxi-SARSA	4/6	68.2	74.2	67.9	**82.44**	86.19
LunarLander-DDQN	8/4	68.4	63.7	72.1	**72.82**	72.91
BipedalWalker-PPO	14/4	56.9	56.4	56.7	**67.99**	69.28
Starcraft-A3C	9/4	94.7	91.8	91.4	**97.36**	86.04

Table 1: Distal explanation model evaluation in 6 benchmark reinforcement learning domains that use different RL algorithms, measuring mean task prediction accuracy in 100 episodes after training. SE-structural equations (trained with LR-linear regression, DT-decision trees, MLP-multi layer perceptrons), *DP*-decision policy tree and DP_n-unconstrained decision policy tree.

Figure 4: Starcraft II Collaborative task scenario: The agent is controlling the leftmost section and the participant controls the right section (divided by the fissure)

actions (*DP* in table 1), other trained until all leaves are pure nodes (DP_n). Results summarised in Table 1 show our model outperforms task prediction of action influence models (with their structural equations trained by either linear regression (LR), decision trees (DT) or multi layer perceptrons (MLP)) in every benchmark, some by a substantial margin.

The benefit gained through unconstrained decision trees (DP_n) does not translate well into an increase in task prediction accuracy. We conclude that for the purpose of using distal models for explanation, a depth limited tree (*DP*) provide an adequate level of accuracy. Moreover, as a depth limited tree is more interpretable to a human, it is more suited for *explainability* and *explanation*.

5 EVALUATION: HUMAN STUDY

We consider human subject experiments to be an integral part of XAI model evaluation and as such conduct a human study with **90** participants. We consider two hypotheses for our empirical evaluation; H1) Distal explanation models leads to a better *understanding* of the agent; and H2) Distal explanation models provide *subjectively* 'better' explanations. Our experiment involves RL agents that complete objectives in 3 distinct scenarios, which are based on the Starcraft II [31] learning environment. We first discuss these scenarios below.

5.1 Scenarios

In addition to the default scenario of the Starcraft II, we developed two additional scenarios as custom maps for the game, that are better suited for explainability. We also release these maps with state and action specifications as test-beds for explainability research.

Adversarial: In this scenario, the agent's objective is to build its base by gathering resources and destroy the enemy's base. The agent can build offensive units (marines) to attack the enemy's base and to defend its own base. This is the default objective in a normal Starcraft II game, but here we only use 4 actions for the purpose of the experiment rewards are given for the number of enemies and buildings destroyed (shown in Figure 3 b) as an action influence graph). During the experiment, the trained RL agent will provide explanations to the participant and the strength of the explanations are evaluated through task prediction.

Rescue: This scenario is a custom map, where the agent's objective is to find a missing unit and bring it back to the base using an aerial vehicle. The agent also has to avoid or destroy enemy units during the rescue and aid the aerial vehicle using an armed unit. The agent has access to 5 actions, the reward is given for the number of missing units saved. The evaluation is done through task prediction as before.

Collaborative Task: The collaborative task is fundamentally different from the previous scenarios, in that the participant has to help the agent to complete the objective. We made this task as a custom map (depicted in Figure 4) where the map is partitioned as the agent and human 'area'. The agent can perform 5 actions in this task, while the human can choose 4 actions to execute. The objective of the task is to build a series of structures that finally leads to the creation of an 'elite' unit, which the human has to transport to a base. The success of the task depends on the participant choosing to execute the action that best support the agent.

5.2 Experiment Design and Methodology

To investigate the two main hypotheses, we use a within-subject design [12] for our experiment. Every participant will be evaluated on the 3 independent variables which are 1) 'no explanations', where only a visual description of the agent behaviour is provided; 2) causal explanations generated with action influence models [22] and 3) our distal explanation model. At a glance, the experiment has 3 phases where participants receive explanations from RL agents, subjectively evaluate the explanation and are then evaluated through task prediction [18] to gauge their understanding of the agent.

Task prediction is an effective measure that can peek into the mental model of an explainee to evaluate how successful the given explanation was in transferring the knowledge from the explainer [18, 25]. In task prediction, the participant is asked the question 'What will the agent do next?'. We use task prediction to evaluate the hypothesis H1) for the Adversarial and Rescue scenarios, and invert the question as to ask 'What would you do next?' in the Collaborative task. We investigate hypothesis H2) by employing the 5-point Likert *explanation satisfaction scale* of Hoffman et al. [18, p.39].

Explanation satisfaction is evaluated after each explanation and also at the end of the experiment which compares explanations of causal and distal models.

Experiment Design: We use *Amazon Mechanical Turk* (Mturk)—a crowd sourcing platform well known for obtaining human-subject data [4]—to conduct the experiments. A web-based interactive interface is used as the medium of interaction.

We record video clips of the agents solving the 3 Starcraft II scenarios that capture the behaviour of the agents. Each scenario has 4 distinct behaviours of the respective agent (around 10 seconds per clip). We first display the ethics approval obtained through a university, and after the participants' consent gather demographic information. Participants can fall into one of the 3 combinations of explanation models and scenarios. For example, a combination can be: Adversarial with no explanations, Rescue with casual explanations and Collaborative with distal explanations.

The first stage of the experiment involves training the participants to identify agents' actions using video clips of the agents performing those actions before the start of each scenario. In the Collaborative scenario, participants are trained to identify the actions they can use instead. After validating that participants can distinguish different actions through a question, the scenario will be presented.

The second stage lets the participants ask explanatory questions (in the form of why/why not *action*), after watching the agent's behaviour through the video clip. Participants can ask any number of questions and we did not control for a minimum number of questions (though we encouraged the participants to ask question by indicating that it can lead to a better score, resulting in bonus payments). After each explanation video, participants are presented with the explanation satisfaction survey. This process is for 4 tasks.

The third stage involves evaluating the participants' 'understanding' of the agent through task prediction. Participants are presented with 4 new videos with different situations, and are asked what action the agent will do next, and can select one of the 4 options (which are 3 actions of the agent plus the option of 'I don't know'). This process is also repeated 4 times. After this stage participant will move to the next scenario with a different explanation model and repeat from Stage 1 to 3. This is done until all the scenarios are encountered by the participant.

In the final stage, the participant is presented with 3 additional explanation videos (of the scenario they did for the no explanation condition), and is presented with causal explanations from action influence models [22] and our distal explanation model *side by side*. We use Hoffman et al. [18, p.39]'s explanation satisfaction scale but this time as a movable slider that subjectively compares the two explanation.

Experimental Conditions:

We ran the experiment with the above mentioned 3 independent variables (the explanation models), which resulted in 3 combinations of explanation model and scenarios. Each combination had 30 participants for a total of 90 participants in the experiment. Each participant is scored on the total number of correct task predictions out of 12 (4 each for each model-scenario combination).

Each experiment ran approximately 50 minutes, and we compensated each participant with 8.5USD (a bonus compensation of 0.5USD was also given to participants for each point above 10). Participants were aged between 23 to 60 (μ = 38.1), and of the 90 participants, 51 were male while 38 were female and 1 was reported as other. Participants reported an average self-rated gaming experience and Starcraft II experience of 2.47 and 1.47 out of 5 (5-point Likert) respectively.

To ensure the quality of data from participants we followed the following methods. We only recruited 'master class' workers with 95% or more approval rate. We controlled for language by only recruiting workers from the United States. We excluded the noisy data of users in 3 ways. First, we tested participants to ensure they had learnt about the scenario by asking them to identify actions shown in several videos. If the participant failed this, the experiment did not proceed. Second, we tracked how much time each participant spent viewing explanations and answering tasks. If this was regularly below a threshold of a few seconds, we omitted that participant from our results. Third, participants were required to explain their task predictions. If this text was gibberish or a 1-2 word response, we omitted that participant from the results.

5.3 Results

We first discuss the results on hypothesis H1), where we investigate whether distal explanation models lead to a better understanding of the agent. We present the null hypotheses as $H_0 : P_N = P_C = P_D$ and the alternate hypothesis as $H_1 : P_D > P_N$ and $H_2 : P_D > P_C$, in which N, C, D corresponds to 'no explanation', causal and distal explanation models. Here, P denotes the proportions of the observed values of correct answers in task predictions by the participants.

To have a transparent presentation of the task prediction results, we summarise the participants' responses in Table 4 as a contingency table. We then perform Pearson's Chi-squared test for the 3 Starcraft II scenarios and obtain the following values: Adversarial (p-value = 0.011, X^2 = 13.00), Rescue (p-value = 0.034, X^2 = 10.40) and Collaborative (p-value = <0.001, X^2 = 35.47). As the Chi-squared test was significant across the 3 scenarios, we investigate the pairwise differences between models using a z-test. We summarise the results in Table 3. From Table 3, considering the proportions (P) between model pairs, we can see that apart from Adversarial and Rescue scenarios for the D - C model pair, distal explanation models have statistically significant results between other combinations. Thus we accept H_1 for every Starcraft II scenario and accept H_2 only for the Collaborative scenario. We illustrate these results as a box-plot in Figure 2. Clearly, the Collaborative scenario poses a much higher challenge to the participants, and results indicate that distal explanations perform better than other models in this task.

Explanation Quality: The second main hypothesis H2) evaluate whether distal explanations can provide *subjectively* better explanations. The corresponding null hypothesis is $H_0 : P_N = P_C = P_D$ and the alternative hypothesis is $H_1 : P_D > P_C$. P in this case P becomes the proportion of the observed values of the Likert scale data (using the survey of Hoffman et al. [18, p.39]), where participants have rated as '5'. We consider 4 explanation quality metrics; 'complete', 'Sufficient', 'Satisfying' and 'Understanding'. As before, we employ Pearson's Chi-squared test to see the significance of the above 4 metrics in the 3 Starcraft II scenarios, and obtain p-values < 0.01 for every condition. As there are significant differences between explanation models on explanation quality, we reject H_0

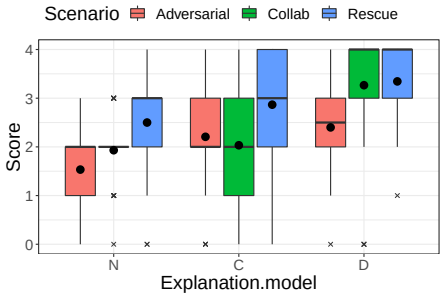

Table 2: Box plot of task prediction scores of the explanation models across the Starcraft II scenarios (means are represented by bold dots).

Mdl-pair	Scenario	X^2	p-value	Proportions
C - N	Adversarial	5.437	0.019	0.53 | 0.38
	Rescue	2.283	0.130	0.71 | 0.62
	Collaborative	0.416	0.518	0.50 | 0.46
D - N.	Adversarial	11.269	**<0.001**	**0.60 | 0.38**
	Rescue	9.931	**0.001**	**0.80 | 0.62**
	Collaborative	31.966	**<0.001**	**0.81 | 0.46**
D - C	Adversarial	1.085	0.297	0.60 | 0.50
	Rescue	2.784	0.095	0.80 | 0.71
	Collaborative	25.511	**<0.001**	**0.81 | 0.50**

Table 3: Pairwise differences with a z-test for proportions for each model pair in the 3 Starcraft II scenarios in task prediction scores, considering the correct response (R column counts in Table 4).

Model	Scenario	R	DK	W
No Exp.	Adversarial	46	7	67
	Rescue	75	11	34
	Collaborative	56	9	47
Causal	Adversarial	64	7	45
	Rescue	86	7	27
	Collaborative	61	7	52
Distal	Adversarial	72	6	42
	Rescue	97	4	19
	Collaborative	98	7	15

Table 4: Contingency table of Starcraft II scenarios and explanation models, R-Right , DK - Don't know, W - Wrong. (high counts in 'R' column are better).

Metric	Scenario	X^2	p-value	Proportions
Complete	Adversarial	3.267	0.070	0.56 | 0.45
	Rescue	0.074	0.785	0.33 | 0.35
	Collaborative	11.428	**<0.001**	**0.40 | 0.20**
Sufficient	Adversarial	6.020	**0.014**	**0.56 | 0.40**
	Rescue	0.018	0.892	0.35 | 0.34
	Collaborative	15.55	**<0.001**	**0.41 | 0.18**
Satisfying	Adversarial	1.085	0.297	0.46 | 0.40
	Rescue	1.528	0.216	0.29 | 0.36
	Collaborative	9.981	**0.001**	**0.42 | 0.23**
Understanding	Adversarial	1.377	0.240	0.46 | 0.39
	Rescue	0.071	0.788	0.35 | 0.37
	Collaborative	15.31	**<0.001**	**0.42 | 0.19**

Table 5: Pairwise differences with a z-test for *explanation quality* metrics in models Distal vs Causal, data where participants rated '5'.

and conduct a pairwise z-test. We summarise the results in Table 5. From the 3 Starcraft II scenarios, only the Collaborative task yield significant results for every explanation quality metric. Thus we accept H_1 only for the Collaborative scenario.

Discussion: The results we obtained for *explanation quality* mirror the results in task prediction. Intuitively this makes sense as participants are more inclined to rate an explanation 'good' if they feel they have a better 'understanding' of the agent. Further investigations are needed to explore why distal explanations perform substantially better in human-agent collaborative tasks.

To investigate whether the knowledge of the Starcraft II game had any impact on task prediction scores, we perform a Pearson's correlation test between task prediction and Starcraft II experience (self-report in a 5-point Likert scale). The obtained values (t = 1.515, p-value = 0.133) indicate that there is no statistically significant correlation between scores and Starcraft II experience. Although our experiment was based on the Starcraft II environment, we used custom maps and scenarios that are different from the game. Thus the results of the correlation test is plausible.

One weakness of our model is the need for a causal graph that is faithful to the problem, in order to learn the *opportunity chains*. For the purpose of this paper, we hand-crafted the causal graphs for Starcraft II scenarios and the 5 RL benchmarks. While our hand-crafted models can be verified easily with data, we acknowledge that it may become infeasible in larger domains. We view generating a causal graph a distinct problem than generating explanations *using* a causal graph. As such we propose this as our immediate future work.

6 CONCLUSION

We introduce *distal explanation* models for model-free reinforcement learning agents that can generate explanations for 'why' and 'why not' questions. These models learn *opportunity chains* (in the form of A enables B and B causes C), and approximate a future action that *enables* due to the current action of the agent. Our motivation comes from insights gained through a human-agent experiment, in which we analysed **240** human explanations. We evaluate our approach in 6 RL benchmarks on *task prediction*. We then undertake a human study with **90** participants to investigate how the distal explanation model perform in task prediction and *explanation quality* metrics in 3 custom Starcraft II scenarios. While results indicate a significantly better performance of distal explanations (against 2 other explanation models) in collaborative situations, further research is needed to understand the impact it has on other types of scenarios. Our immediate future work involves generating causal models at the training time of the agent through causal discovery.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. ACM, 582.

[2] Dan Amir and Ofra Amir. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. (2016). arXiv:arXiv:1606.01540

[4] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[5] Ruth M. J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the 28th International Joint Conferences on Artificial Intelligence*. 6276–6282.

[6] David Chapman and Leslie Pack Kaelbling. 1991. Input Generalization in Delayed Reinforcement Learning: An Algorithm and Performance Comparisons.. In *IJCAI*, Vol. 91. Citeseer, 726–731.

[7] Kathy Charmaz and Linda Liska Belgrave. 2007. Grounded theory. *The Blackwell encyclopedia of sociology* (2007).

[8] Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

[9] Francisco Elizalde and Luis Enrique Sucar. 2009. Expert Evaluation of Probabilistic Explanations.. In *ExaCt*. 1–12.

[10] Francisco Elizalde, L Enrique Sucar, Manuel Luque, J Diez, and Alberto Reyes. 2008. Policy explanation in factored Markov decision processes. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM 2008)*. 97–104.

[11] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. 2017. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In *Proceedings of the 5th International Conference on Human Agent Interaction*. ACM, 97–101.

[12] Anthony G Greenwald. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83, 2 (1976), 314.

[13] Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science* 56, 4 (2005), 889–911.

[14] Bradley Hayes and Julie A Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. ACM, 303–312.

[15] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[16] Denis J Hilton, John McClure, and Robbie M Sutton. 2010. Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology* 40, 3 (2010), 383–400.

[17] Denis J Hilton and John L McClure. 2007. The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking*. Routledge, 56–72.

[18] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608* (2018).

[19] Jennifer Hornsby. 1993. Agency and causal explanation. (1993).

[20] Omar Zia Khan, Pascal Poupart, and James P Black. 2009. Minimal Sufficient Explanations for Factored Markov Decision Processes.. In *ICAPS*.

[21] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Toward Robust Policy Summarization. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2081–2083.

[22] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable Reinforcement Learning Through a Causal Lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[23] John McClure, Denis J Hilton, and Robbie M Sutton. 2007. Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European journal of social psychology* 37, 5 (2007), 879–901.

[24] Tim Miller. 2018. Contrastive Explanation: A Structural-Model Approach. *arXiv preprint arXiv:1811.03163* (2018).

[25] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[27] Aaron M Roth, Nicholay Topin, Pooyan Jamshidi, and Manuela Veloso. 2019. Conservative Q-Improvement: Reinforcement Learning for an Interpretable Decision-Tree Policy. *arXiv preprint arXiv:1907.01180* (2019).

[28] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[29] Steven Sloman. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

[30] Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerincx. 2018. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706* (2018).

[31] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).

[32] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.

[33] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 109–116.