Policy Graphs and Intention: answering 'why' and 'how' from a telic perspective

Victor Gimenez-Abalos*
Barcelona Supercomputing Center
Barcelona, Spain
victor.gimenez@bsc.es

Sergio Alvarez-Napagao*
Universitat Politécnica de Catalunya
Barcelona Supercomputing Center
Barcelona, Spain
salvarez@cs.upc.edu

Adrian Tormos
Barcelona Supercomputing Center
Barcelona, Spain
adrian.tormos@bsc.es

Ulises Cortés
Universitat Politècnica de Catalunya
Barcelona Supercomputing Center
Barcelona, Spain
ulises.cortes@bsc.es

Javier Vázquez-Salceda
Universitat Politécnica de Catalunya
Barcelona, Spain
jvazquez@cs.upc.edu

ABSTRACT

Agents are a special kind of AI-based software in that they interact in complex environments and have increased potential for emergent behaviour. Explaining such behaviour is key to deploying trustworthy AI, but the increasing complexity and opaque nature of many agent implementations makes this hard. In this work, we reuse the Policy Graphs method –which can be used to explain opaque agent behaviour– and enhance it to query it with hypotheses of desirable situations. These hypotheses are used to compute a numerical value to examine agent intentions at any particular moment, as a function of how likely the agent is to bring about a hypothesised desirable situation. We emphasise the relevance of how this approach has full epistemic traceability, and each belief used by the algorithms providing answers is backed by specific facts from its construction process. We show the numeric approach provides a robust and intuitive way to provide telic explainability (explaining current actions from the perspective of bringing about situations), and allows to evaluate the interpretability of behaviour of the agent based on the explanations; and it opens the door to explainability that is useful not only to the human, but to an agent.

KEYWORDS

XAI; intentions; post-hoc explainability; Agent Explainability; Telic Explanations; interpretability; reliability; Explainable Agency

ACM Reference Format:
Victor Gimenez-Abalos*, Sergio Alvarez-Napagao*, Adrian Tormos, Ulises Cortés, and Javier Vázquez-Salceda. 2025. Policy Graphs and Intention: answering 'why' and 'how' from a telic perspective. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

* Corresponding authors.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Among the tasks within the purview of Artificial Intelligence (AI), the issue of solving problems without giving explicit instructions on *how* to solve them is pervasive. However, precisely because of the definition of such a task, the result is an *artefact* that, unless explicitly designed to be transparent, is often not interpretable or, hence, trustworthy [29, 64]. This is where the field of *Explainable Artificial Intelligence* (*XAI*) shines through.

Explanations can be viewed as a communicative exercise between source (*i.e.* the model or one of its components) and receiver (*i.e.* the explainee) that describes the relevant context or the causes surrounding some facts [28, 39, 63], which in the context of AI is often related to its final or intermediary outputs or decisions.

Although any such communicative act can be considered an explanation, some explanations are better than others. Considering explanations as a cooperative communication, Herbert Paul Grice would rate explanations according to four utilities [20]: how interpretable the content is to the receiver (*manner*), how truthful it is (*quality*), how concise it is (*quantity*), and how relevant it is to the context of the communicative act (*relation*).

In this paper, we focus on the first two. On one hand, we highlight the relevance of reliability: whether the explanation given by the model is factually correct, which is dependent solely on the sender though it is sometimes sacrificed in pursuit of interpretability. On the other hand, we consider interpretability as how much of the explained behaviour can the receiver comprehend and leverage, which is dependent on the receiver. These two separate optimisation objectives tend to be in conflict (*e.g.* the most reliable explanation would involve a detailed breakdown of its code, while the most interpretable explanation might be a simplified and potentially misleading description of its behaviour).

The trade-off ought to be considered pragmatically: *What is explainability used for?* Regardless of context and the nature of the source of explanations, humans find explanations helpful for several aspects, including [1]: for the sender to *justify* behaviours so that the receiver understands it and to hold accountability, responsibility and transparency; for the receiver to *control* and correct the sender's model via locating flaws and vulnerabilities or to debug; for the sender to *improve* based on feedback from the receiver, such as

inspecting nonsensical behaviours and increasing rationality; and for the receiver to *discover* or learn what knowledge the sender has, and how it leverages it to their advantages.

Any desirable *XAI* algorithm is tackling at least one of these objectives [1, 29, 39] while holding some notions (often implicit) of the desirability of explanations related to some of Grice's maxims. When performing explanations over models which can be easily accessed, this task is already complex enough. But as models are becoming increasingly opaque, this task becomes more complex [6, 22], sometimes becoming unachievable. We, as a community, need better tools to tackle this problem [31]. This is particularly the case for autonomous agents [15] that interact in an environment. It is tough to understand an agent's purpose or assumed intentions (*e.g.* a cleaning bot that stops cleaning and goes away), especially if one has no access to its model or it is opaque. In these cases, obtaining explanations becomes an exercise in anthropomorphism, where a human interpreter attributes behaviours (based on what a human would do, as shown by [24]) in a qualitative analysis (*e.g.* the bot has low battery, it goes away to recharge) that may be inaccurate and risks self-deception and harm [48, 62]. However, it is through these mechanisms that humans have evolved to interpret and coordinate with other humans (and even phenomena) through time, marking it as a viable (if dangerous) strategy. If these mechanisms could be turned into quantitative, verifiable, and reliable explanations, they could be used to increase the trustworthiness of AI-based systems by allowing the *explainee* to compare explanations provided and to be aware of their quality and manner.

In this paper, we introduce the notion of intention into *Policy Graph* (*PG*), a *XAI* technique applied to agents in order to explain their behaviour by extending previous attempts [11, 23, 30, 57]. The original technique is based on taking observations of opaque agents to build a symbolic representation of states and actions in a frequentist manner, compiling them into beliefs of policy and state-transition behaviour based on reliable evidence. To this technique, we introduce human-hypothesised desires as annotations into this representation, leveraging the fact that humans can both easily hypothesise behavioural goals based on information and that explanations involving those goals would be interpretable to them – as they have come up with them in the first place. We use these annotations to produce agent intentions as a function of how likely the agent is to reach the annotated goals from a set state, and use these to produce telic explanations, involving the ends rather than the means of the agent. We introduce two metrics that evaluate reliability and interpretability of behaviour, tuned by a hyperparameter that explicitly controls the trade-off in the provided explanations.

This information can be used for *XAI*, with example algorithms answering "*Why* did you take a certain action" as a function of how it helps bring about a goal compared to other possible actions, "*What* do you plan to do" by the intentions –which can be visualised in real time for quick examination of the agent–, and "*How* do you plan to achieve something" as the beliefs that support that fulfilling a desire is possible by following a course of action. We believe these questions open the door to *justify*, and *discover* agent behaviour, and opportunities to *control* and *improve* it, and enable further downstream tasks like collaboration and/or competition in multi-agent systems, human collaboration, and especially auditing of such systems [23, 49, 55].

2 BACKGROUND

Our focus in this paper is to tackle explainability for *opaque* and *unknown* agents: all the prospective explainee can do is observe its behaviour. Much like humans do when explaining other humans, we frame the problem in the context of no knowledge of the internals of an agent, nor complete observability of the context in which the agent takes its decisions. However, we will assume that the explainee may have access to a (potentially incomplete) initial notion of what the desirable behaviour should be in terms of what is needed in order to control, improve or justify the actions of the agent [1, 32], from an explainee point of view.

When considering the categories of explainability in which the proposed technique falls in the context of explainable agency (and *Reinforcement Learning* (*RL*) as a subset), current taxonomies [1, 2, 4, 38, 43, 65] offer different categories. Our method is *post-hoc* [1, 43]: explainability applied to agents after its creation, without changing its internal structure to be more transparent. It is *model-agnostic* [1, 43], as we require no knowledge of the internals – neither architecture nor any knowledge of reward function [21] or task decomposition [7, 59] – which is a less tackled problem compared to model-specific [3, 8, 14, 27, 34, 46, 60, 61]. On the context of the scope of explanation [1, 13, 43], we focus on both the *global* and *local* aspect: decisions are taken in states, but include future and global model behaviour to justify those decisions. Finally, in the context of *RL*, a distinction is made between what part of the internals of an *RL* agent is explained [38]. Albeit we remain agnostic to the internals of an agent, within the limitations of each category, our method best matches *Policy-level* explainability as our method focuses on long-term behaviour – rather than single-actions – and is agnostic to how or whether the agent learnt its behaviour.

In addition to these categories, we emphasise that the kind of explanation motivating this method goes beyond the original *PG* of atomic action selection. Instead, behaviour predictability is important for producing relevant explanations. As shown in Malle's research [35, 37], explanations related to understandable end-goals, desires, or rewards are preferable to ones linked exclusively on means. For this purpose, *Theory of Mind* (*ToM*) may become a powerful ally [18, 19, 25], focusing on easing the task of behaviour predictability through the concept of intentions [10, 19, 36, 42].

Intentions are mental states different from others such as beliefs, desires, knowledge or emotions. In more *agentic* frameworks, an intention consists on a state of affairs that will be the aim of the agent and to which it commits [9], while in humanistic ones, intentions are the result of both desiring and believing to have a course of action that brings about a state of affairs [35, 37]. These definitions support our decisions to model intentions in *PG*s (see § 3.1).

However, especially when dealing with opaque agents, attributing intent can be dangerous – especially since it can be noted that not all agents may have explicit intentions. Therefore, this attribution may not be completely right from a formal perspective [63], but the fact it is practical and beneficial to do so underpins the motivation why we do this – humans do this attribution process constantly to explain affairs. Instead, the *burden of attribution* ought to be considered from the point of view of whether it is useful, reliable, and interpretable, much like with other communicative acts.

3 METHODOLOGY

*PG*s are a probabilistic graphical model that compiles two pieces of knowledge: the agent policy – as $P(a|s)$ – and a world model [16, 17, 45] – as $P(s'|a, s)$. These two pieces of information, together with the assumption of the Markov property –that is, that the probabilities of actions and transitions are independent of previous states given you know the current one– allow for computing the probability of any sequence of actions and states, starting from one.

These quantities, when actions and states considered are finite, allow for a graph-like representation, where nodes are states and edges are annotated with actions and encode the transition probability between states. In previous work [11, 23, 30] there are proposals on how to collapse continuous states (or actions) to discrete versions with a discretiser, whose influence is analysed in those works. The discretiser converts the state into predicates, which we leverage later on for explainability-generation purposes, as well as for the introduction of intentions. The *PG* is then built in a frequentist fashion: taking observations of agent behaviour, converting states and actions into discrete versions, and computing the required probabilities by counting occurrences. In addition to these quantities, in this work we also add to the representation $P(s)$ – as the number of times out of the total that we visited a state. Note that in this process, no access to the policy or agent internals is required, and the origin of each quantity can be traced into empirical evidence.

Following this process, a *PG* is obtained, and previous work allows for different kinds of explainability to be applied. In this work, we extend this representation into an Intention-Aware *PG*.

3.1 Explainability with Desires and Intentions

Most explainability algorithms in literature speak of causal relationship, correlation, or *relevance* between the model's output and some input quality [33, 44, 52]. However, when translated into agent behaviour, the responses can be quite different from the ones we expect from, for example, humans. For example, when asking a chef why they put a pot on a hob, they will not reply in terms of the pot being full of water, or the hob being unused.

Explanations involving human intent are often teleological. They summarise behaviour by referencing a root cause: the ends by which a course of action is chosen. In many cases, these teleological explanations encompass the realms of morals, ethics and politics [26, 63], but the actual intention acts as the main predictor of the existence of abstract mental states. When asking a human cook why they put a cooking pot on the hob, they would reply: *Because I want pasta carbonara, for which I need to cook the pasta...* eventually arriving at *I can achieve it by putting the available water-filled pot on the unused hob*. Depending on the *beliefs* of the explainer about the necessities of the explainee, the explanation would be cut short, taking the slice of the whole trace of reasons that they think is relevant for the explainee. The explainer cannot reasonably constrain itself to the lowest level alone [61].

To achieve explainability of this sort, our first observation is that humans do this already: when analysing a (reasonably well-performing) agent's behaviour in a domain, humans tend to anthropomorphism [24, 48, 62], establishing logical inferences from a teleological perspective [51, 63]. Humans attribute intentionality to the agent. This is especially the case for most toy environments (*e.g.* games) of which the human observer has some knowledge of how to solve and thus is expecting certain behaviours. This extends to experts observing agents' behaviour in their domains [40, 54]. Explanations in terms of these attributed intentionalities are, by virtue of being hypothesised by a human, more understandable for humans. However, as is now, this is a dangerous affair. Such attributions are subject to anecdotal fallacy when observing a low number of interactions.

Instead, we aim to leverage these attributions by making them testable, and presenting metrics in terms of how reliable it is to perform that attribution, and how much of the agent's behaviour can be explained by the attribution. These attributions can then be used to enable the agent to answer the *what*, *why*, and *how* questions in a manner not dissimilar to how a human would. This is done through the introduction of *agent desires*, which can be modelled in diverse ways, and *agent intentions*, that is, the desires we expect the agent to accomplish (soon) as allowed by the environment [9].

3.1.1 Desires and Intentions. In this work, we refer to the hypotheses over expected behaviour as *desires*. This desire may or may not express itself (or not do so frequently) in the behaviour of the agent, and thus they require verification. If a desire truly expresses itself, it is often due to the design concerns through which the agent was created, be that some particular rule in the system, the design of a reward function, or a statistical bias in the data it trained on.

Pragmatically, defining a desire requires understanding when it is fulfilled. For this work, we limit to desires of the form of a perform goal [58], to execute an action in a state where some qualities hold (*e.g.* for a cleaning bot to dock into a charger when the house is clean and it is near the charging port), as they are the most useful for our use-case, but it is important to note that other types (*e.g.* achievement goals) may be modelled equally.

A desire d can be defined as a tuple $\langle S_d, a_d \rangle$: a discrete state region $S_d = \{s \in S | s \vdash d\}$ where $s \vdash d$ means that the state satisfies the desire's condition, and the action a_d that would be desirable in such state region. As the explainees themselves provide this characterisation, they are expected to understand it when it becomes the *finality* of explaining behaviour. For the cleaning bot example, the desire to charge when having low battery could be modelled with a_d being docking and S_d any state with low-battery and being close to the charging port.

Calculating relevant information over these desires is trivial under the probabilistic description of a *PG*. *How likely are you to find yourself in a state where you can fulfil your desire by performing the action?* can be computed as the desire state region probability $P(s \in S_d) = \sum_{s \in S_d} P(s))$. *How likely are you to perform your desirable action when you are in the state region?* can also be computed as $P(a_d | s \in S_d) = \sum_{s \in S_d} P(a_d|s) * P(s)/P(s \in S_d)$.

However, this view is rather myopic: no information appears available outside states in an S_d. Most states in a typical problem do not manifest the specific conditions for immediately fulfilling a desire. To solve this, we extend desire information backward through the state transitions, in what we refer to as intentions.

An agent's intention to fulfil a desire exists if it can be fulfilled (given by world dynamics and its understanding), and the agent commits to doing so [9]. The empirical observations of the agent's behaviour capture both requirements: the world model captures

the possibility of a desire being fulfilled, and the policy captures the agent's committing or working to achieve them.

Loosely defined, given a desire $\langle S_d, a_d \rangle$, the agent's intention of fulfilling it in a state s ($I_d(s)$) can be measured as the probability that the agent will attain the desire from s. Informally, it is the sum of probabilities of all possible state-action sequences which end up bringing about d: a sequence ending in $[s_i, a_d]$, with $s_i \in S_d$.

Let $\mathcal{T}(s, d)$ be the (potentially infinite) set of trajectories (sequences of states and actions) originating from s and arriving at any $s' \in S_d$. The intention of the desire in state s can be computed with the *PG* information as:

$$I_d(s) = \sum_{seq \in \mathcal{T}(s,d)} P(a_d | last\_state(seq)) P(seq)$$

where $P(seq)$ is the probability of seeing a sequence of states and actions seq as computed by the *PG*:

$$P(seq) = \Pi_{t=1}^{|seq|} P(s^{t+1} | a^t, s^t) P(a^t | s^t)$$

To succinctly deal with infinitely-looping paths, we compute intention backwards: starting from S_d and recursively propagating intention updates to parent states. A stopping criterion ϵ is introduced to stop the propagation of intentions below a certain probability. Algorithms 1 and 2 illustrate the process. We note that desire propagation is stopped from crossing through transitions that would fulfil them, as not doing so would compute the 'expected number of times a desire will be fulfilled' instead of a probability.

Algorithm 1 Register a Desire into a PG and propagate intentions

Require: $d = \langle S_d, a_d \rangle, PG$
  for $s \in PG$ **do**
    $I_d(s) \leftarrow 0$
  end for
  for $s \in S_d$ **do**
    $increment \leftarrow P(a_d|s)$
    `Propagate_intention`$(s, d, PG, increment)$
  end for

Algorithm 2 Propagate intentions to node s.

  procedure PROPAGATE_INTENTION(s, d, PG, inc)
    $I_d(s) \leftarrow I_d(s) + inc$
    for $p \in \{p \in PG | P(S' = s | S = p) \neq 0\}$ **do** ▷ All parents of s
      if $p \notin S_d$ **then** ▷ P cannot fulfil the desire
        $propagable\_inc \leftarrow P(S' = s | S = p) * inc$
      else ▷ P could fulfill the desire by doing a_d, ignore a_d
        $propagable\_inc \leftarrow P(S' = s, A \neq a_d | S = p) * inc$
      end if
      if $propagable\_inc \geq \epsilon$ **then** ▷ Stop criterion
        `Propagate_intention`$(p, d, PG, propagable\_inc)$
      end if
    end for
  end procedure

$I_d(s)$ is a useful tool for answering complex queries: *What do you intend to do in state s?*, can be replied with all desires with an $I_d(s)$ over a certain threshold; *How come you believe you can do d*, which can be replied with empirical evidence of how one can bring about d, using descendants $I_d(s')$ to prospect the graph; *Why did you take action a at state s?*, can be replied to in terms of desires as: *I have the desire d, which I can bring about from state s, and by performing action a either I am closer to achieving it, or there is a chance I will increase my odds of doing so.* The algorithms for replying to these queries can be found in § 3.1.2.

The intention value is easily interpretable, as it is the probability that some desire will be brought about by the agent from a state. However, the lower the intention, the more uncertain its fulfilment, and the continuous property of intentions makes it so that a user may convince themselves of wrong information by vastly overestimating a probability. For this, we propose to restrict intention attribution to intentions above a parameter: the *commitment threshold* $0 < C \leq 1$. This parameter specifies the scepticism of an explainee: at which minimum probability the explainee is willing to believe the agent will try to fulfil a desire. For safety and reliability purposes, any $I_d(s) < C$ is to be disregarded and not used to produce explanations, whereas, for any state s such that $I_d(s) \geq C$, the agent can be said to have (at least some) intention to fulfil d. We refer to the set of states with a desire d attributed as $S(I_d)$.

This *commitment threshold* is a parameter directly related to the reliability-interpretability trade-off. With low C, more intention is attributed and thus more explanations can be provided (and thus behaviour is apparently more interpretable) at the cost of attributing courses of action that are improbable (thus providing unreliable explanations). With high C, the likelihood of intentions being fulfilled increases as does reliability of explanations including them, at the cost of leaving more behaviour unexplained. Both of these quantities can be measured given a commitment threshold.

Formally, we call $P(s \in S(I_d))$ the **attributed intention probability**, *i.e.* the probability that, at any point of observation, the agent is attributed the intention: $P(s \in S(I_d)) = \sum_{s \in S(I_d)} P(s)$. This metric estimates interpretability of behaviour: as intention attribution is more likely, it is easier to answer to why it is acting. Conversely, we call **expected intention** to the probability that, once attributed, an intention will be fulfilled:

$$\mathbb{E}_{s \in S(I_d)}(I_d(s)) = \sum_{s \in S(I_d)} I_d(s) * \frac{P(s)}{P(s \in S(I_d))}$$

This second metric represents reliability of explanations: the higher the value, the more likely an attributed intention comes to pass. Although perfection in both these metrics seem an absolute requirement for explainability, real scenarios leave this option likely out of reach. On one hand, for a sufficiently low C and enough desires considered, it is likely possible to reach perfect *intention probability* (*i.e.* always being able to attribute *why*), but at the cost of being wrong several times. On the other hand, even with a high C value and a perfectly rational agent, in stochastic environments it is possible (even likely) that an agent with an intention fails to achieve it due to unexpected environment changes. As such, we believe crucial to not only measure, but also let the trade-off be explicit and in control of the explainee.

Finally, beside these desires, we introduce the *Any* desire to aggregate all. This desire is the aggregate of all other desires, corresponding to a fake intention $I_{Any}(s) = max_d I_d(s)$. In turn, the previous metrics computed for *Any* desire aggregate the reliability and interpretability of explanations overall.

3.1.2 Using Intentions in Explanation Algorithms. Previous work argues that the form of any explanation must account for its function as an answer to a why question [53], highlighting the intrinsic causal nature of explanations. However, the intent behind a 'why' question can be varied, and humans tend to have greater context to understand which kind of answer is expected of them. In the folk-conceptual theory of behaviour explanation, one can categorise between explanations provided for unintentional and intentional behaviour [35, 37]. Where most previous work [11, 23, 30] focuses on answering *why* queries as a function of beliefs about the current state (of which the agent has seemingly no agency over when the question is asked) and thus are of a more unintentional optic. Instead, we shift the focus to answering in the form of intentions and desires for intentional behaviour explanation: asking questions such as "*What for?*" as a more specific version of "*Why?*".

Intentional behaviour-related questions can be categorised into three modes [35]: *Reasons Explanation* (*RE*), which are by far the most common, examine the causality of action based on what intentions the agent has and how an action furthers them; *Causal History of Reasons* (*CHR*), which assume that the intent of the action is apparent and rather focus on the reason behind it being desirable; and *Enabling factors* (*EF*), which justify why an action that is apparently desirable was successful (generally when its success was perceived to be improbable or difficult). To cover these types of modes, we propose some queries, together with a shared vocabulary between questions and answers that allow to make question chains. The queries and answers form are the following[1]:

- "*What* do you intend to do in state s?": It answers a part of *RE* (possible intent of actions performed in a state). The reply is in terms of desires d as well as intention values (*i.e.* probabilities $I_d(s)$).
- "*Why* do you do action a in state s?": This responds to the action-side of a Reasons Explanations, by examining how a could increase the odds of a desire d being fulfilled by bringing you to a different state s'. It could also be stated as a *what for* question, but we keep the *why* form for the purpose of having different keyword identifiers for the queries.
- "*How* do you plan to fulfil intention I_d from s?": this follow up question can be viewed as answering *EF*, by making *PG* beliefs explicit over a course of action bringing about a desire d. When chained after the previous question, one can question why being in a state s' increases intention: showing the possible future paths from s' that bring about d.

The first question (*What*) is trivially answered within the framework: given a state s, it returns all attributed intentions: $\{I_d(s) \geq C\}$. We use the commitment threshold to ensure the reliability of the answer to this question is within the desired by the explainee, discarding low-valued intentions.

To answer the second question (*Why*), let us consider: if no intention exists in s, then the action is not intentional –that is, caused by an intent– from the perspective of the *PG* and no answer should be returned. Else, each attributed intent is a candidate, to be tested independently. If the intent is the cause of an action, this action must result in an increase of intent: this can happen in two ways. Either the action results in an expected increased intent, or the action is a gamble that can result in an increased intent. The former can be expressed as:

$$\mathbb{E}_{P(s'|a,s)}(I_d(s')) - I_d(s) = \sum_{s'} P(s'|a,s) * I_d(s') - I_d(s)$$

The latter can be expressed in terms of the probability of an increase $P(I_d(s') \geq I_d(s)|s,a)$ and expected *positive* increase:

$$\mathbb{E}_{P(s'|a,s,I_d(s') \geq I_d(s))} I_d(s')$$

Instead of either aggregation, a histogram of $P(I_d(s') - I_d(s)|s,a)$ could be used for visual analysis. It is worth noting that contrastive explanations can be built by using two *why* questions over different actions, and examining the different replies. With answers possibly irrational (*e.g.* in case of *RL* agents with suboptimal policies or vestigial exploration behaviour), this opens the possibility of using explanations to *improve* or *control* agent behaviour.

Finally, for the third question (*How*) which examines by extension *Why* the agent believes a desire can be achieved from state s, the *PG* models are used to plot future trajectories in two possible ways, in Algorithms 3 and 4.

The former is a simpler, optimistic explanation: assuming the world and agent behave in optimal ways, starting from s choose s', a that maximise $I_d(s')$ restricted to $P(s', a|s) > 0$ (*i.e.* being a valid successor). Given that $I_d(s)$ is computed as an expectancy of successor intentions, $I_d(s') \geq I_d(s)$ is guaranteed. This process is iterated until the desire is fulfilled, and a sequence of actions and states is returned. Given the predicate-like nature of the representation, the response can be given concisely by returning predicate changes (as opposed to full state descriptions). However, this algorithm gives a plausible path but does not account for setbacks or world stochasticity. Algorithm 4 complements the answer by considering sampling random state successors from $P(s', a|s)$ iteratively, recording multiple possible sequences. This results in the possibility of a sequence arriving in a state where the intention is no longer attributed (*i.e.* falls below C). As such, this algorithm returns both belief-evidence justifying the value of $I_d(s)$.

Algorithm 3 How do you plan to fulfill d from s?

```
procedure HOW(d, s, PG)
    current ← s
    if s ⊢ d then                          ▷ State can fulfill desire
        return a_d                         ▷ return action that fulfills the desire
    end if
    s' ← argmax_{s',a∈Succ(s)} I_d(s')
    return Concat(a, s', how(d, s', PG))
end procedure
```

There is a limitation to these questions: if a certain course of action was never performed during the *PG* construction, it is not possible to reason or perform explanations with these algorithms without further assumptions. We discuss this further in § 5.

Finally, a last method of *XAI* is considered based on the needs for quick explainability. As explainability is useful to predict behaviour and coordinate with agents, a real-time visual intention tracking graph can be plotted using $I_d(s)$ in real time. Since I_d can

[1]Notice how no question tackles *CHR* yet as we argue it has less priority in this setup: reasons about why a desire d hypothesised by the explainee ought to make sense to the explainee; and thus remains future work.

Algorithm 4 Stochastic how do you plan to fulfill d from s?

```
procedure HOW_STOCHASTIC(d, s, C, PG)
    current ← s
    if s ⊢ d then                          ▷ State can fulfill desire
        return a_d, Success ▷ return action that fulfills the desire
    end if
    if I_d(s') < C then                    ▷ Intention is no longer attributed
        return Failure
    end if
    s', a ~ P(s', a|s)
    return cat(a, s', how_stochastic(d, s', C, PG))
end procedure
```

be precomputed, its algorithmic cost is tied to the discretisation algorithm only. The resulting plots are easily interpretable at a glance, showing what the current intents of the agent is, allowing to easily predict future agent behaviour intuitively.

4 EXPERIMENTS

In this section, we present empirical results for the application of the methodology and metrics to a concrete use case: the Overcooked-AI environment [5], which allows for different environmental layouts, with different characteristics and challenges. This is the proposed experimentation methodology:

(1) In § 4.1 we explain the environment, layouts and agent-training methods used. For each layout, we train specialised agents from scratch. Different combinations of agents are tested. The performance of the agents is computed according to the environment rewards.
(2) In § 4.2 a discretiser tested in the literature [11, 57] is used to build the *PG* from a dataset built through observations of the trained agents. A set of desires that seem relevant for the Overcooked-AI scenario is created and I_ds are computed.
(3) In § 4.3 intention metrics are computed for the policy graphs, showcasing which insights of agent behaviour can be gleaned from the metrics alone. The reliability-interpretability trade-off is evaluated through ROC-like curve comparing the two metrics as C is changed.
(4) Finally, in § 4.4 some example outputs of the explainability algorithms are shown.

All experiments are done on the Overcooked-AI environment. Training code has been developed using Pantheon-RL [47]. The library for producing the policy graphs is *pgeon* [56][2], being developed by the authors and other contributors.

4.1 Environment, Layouts and Agents used

Overcooked is a simple multi-agent environment in which agents collaborate with the purpose of serving dishes. In the scope of this paper, this is restricted to a 2-agent environment, in which agents must cook soup. Agent action-space consists of four directional displacements (`Up, Down, Left, Right`), a non-action (`Stay`), and an action to `Interact`. Displacement only works in walkable spaces (salmon-coloured tiles), and only if the tile is unoccupied: agents cannot occupy the same tile. Interact permits to grab items

[2]https://github.com/HPAI-BSC/pgeon

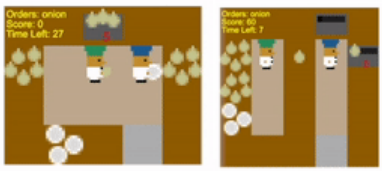


Figure 1: Visualisation of the Simple and Random 0 layouts.

	PPO1, PPO2	HAg, HPPO	RAg, PPO2
simple	387.9 (25.3)	251.3(31.6)	21.6 (16.7)
random0	395.0 (54.4)	108.0 (46.5)	7.6 (6.0)

Table 1: Reward mean and Std Dev of the agent pairs.

if the agent faces in the direction of an item depot (in our case, onions or dishes), drop them on an empty counter (non-walkable, unoccupied brown tile), interact carried items with a pot (with diverse effects), or deliver a cooked dish in the service tile (non-walkable grey tile). Figure 1 displays example layouts of tiles, the ones used for experimentation in this paper.

The game-loop is: agents must interact with onion depots to grab onions, then interact with pots to put them in. Once three onions are in a pot, it waits for 20 game ticks. After it ends, the pot can be interacted with a dish to scoop onion soup. Lastly, onion soup can be delivered in the service tile, providing score to both agents irrespective of which one delivered. Although tricks can be done to smooth the reward function in *RL* agent training, the only action that provides score is the last one.

The two layouts used for this paper are commonly used in the literature. Simple is cramped, and agent collision is common. Random 0 is an asymmetric scenario where cooperation is not just optimal but necessary, as allowances available are different and agents require passing items over a counter.

The agents examined in this paper are mainly *RL* ones. Agents are trained in pairs, with the reward being the environment's default without additional intermediary incentives. The training routines for *RL* were not left to converge into an optimal policy, and thus still exhibit some irrational or erratic behaviour at times. This highlights in the results that the technique works not only for perfectly optimal and rational agents. The agent pairs examined are the following:

- *Pair A* (PPO1 (Blue), PPO2 (Green)): two agents trained from scratch with *Proximal Policy Optimisation* (*PPO*)[50]. These agents were used to validate *PG*s in previous work [12].
- *Pair B* (HPPO (Blue), HAg (Green)): A *PPO* agent trained to collaborate with a human-like agent created via imitation learning on human play-data. The pair was used in previous work [5, 57]. Note that some behaviours learnt by the *PPO* agent (HPPO) are suboptimal given the lack of co-adaption.
- *Random baseline* (RAg (Blue), PPO2 (Green)): similar to *Pair A* but PPO1 is substituted by an agent that samples actions from a uniform probability distribution (all actions have probability 20% regardless of the state). This agent pair is used as a baseline for comparison with the other two pairs.

For each layout, the agents were trained from scratch, so there are a total of nine different agents, and a total of six arrangements (with RAg being in both layouts). The performance evaluation can be seen in Table 1, computed as means and standard deviations of 500 episodes in each layout per agent pair.

Variable	Domain
held	O, D, S, ∅
pot_state$_i \forall i \in [n_pots]$	Empty, Cooking Waiting, Finished
item_pos$_i \forall i \in \{0, D, S, Pot, service\}$	↑, ↓, ←, →, I, S
held_partner	O, D, S, ∅
partner_zone	N, NE, E, SE, S, SW, W, NW

Table 2: Variables and domains of the discretiser.

4.2 Discretiser and Desires

For the construction of the *PG*s, we use a discretiser from literature [11, 57] summarised in Table 2. It consists on five kinds of predicates: the contents of what the observed agent is carrying (held, either **O**nion, **D**ish, **S**oup, or nothing), the state of (each of) the pot(s) (pot_state, Empty: no onions, Cooking: 1-2 onions, Waiting: 3 onions, or Finished: soup), the action that would bring you closest to each interesting item: onion, dish, pots, soup, or service area (item_pos), what the partner agent is carrying (held_partner) and the location of the partner relative to the agent (partner_zone). The *PG*s have been generated from observing 1500 episodes, with up to 400 steps per episode.

With these predicates, we do the exercise as potential explainees watching agent behaviour and come up with two initial types of desires: agents putting onions in the pot, and agents delivering soup at the service area. We remark how the former, while being conductive to obtaining reward down the line, is not explicitly rewarded by the training methods used in § 4.1. In addition, in multiple-pot layouts, there may be differences in how onions are placed in pots based on how many there are in it: intuitively, topping up a pot is better, as cooking time is an overhead which can be parallelised with filling up a second pot. The discretiser allows us to distinguish between putting an onion in an Empty or a Cooking pot, so we register the following desires:

(1) To start cooking: $S_{start\_cooking}$ are states where the agent holds Onion, and is in interact position with a Empty pot; $a_{start\_cooking}$ is Interact. In case of multiple pots, one instance of each desire is created per pot.
(2) To cook: Same as above, but with a Cooking pot.
(3) To service: $S_{service}$ are states where the agent holds soup and has the service in Interact position; $a_{service}$ is Interact.

More desires could have also been hypothesised, such as picking onions when the pot is not full, or a dish when it is waiting. There is no drawback to using those, and they could perfectly be explored too, increasing behaviour interpretability –as more states are attributed intentions– without affecting reliability.

4.3 Intention Metrics

After registering the desires, we compute the **attributed intention probability** and **expected intention** metrics for each of the layouts and agent pairs. Figure 2 show these metrics for the two agent pairs (and random) in the same layouts (Simple and Random 0) and a C of 50%. This information can be used to gauge how likely the method is for providing satisfying explanations to the explainee. To study the C-thresholds, we used Figure 3 to visualise the trade-off between interpretability and reliability for all layouts and agents.

Interact(0.82)	Right(0.89)	Down(1.0)	Interact
held(S) pot_state (Pot$_0$;Empty)	item_pos(O;I) item_pos (Pot$_0$; ←) item_pos (Service;↓) item_pos(S;↓)	item_pos(O;↑) item_pos (Service;I) item_pos(S;→) pot_state (Pot$_0$;Cooking)	
held(D) pot_state (Pot$_0$;Finished)	item_pos(O;→) item_pos (Pot$_0$;I) item_pos (Service;→) item_pos(S;→)	item_pos(O;I) item_pos (Service;↓) item_pos(S;↓) pot_state (POT$_0$;Empty)	

Table 3: Deterministic answer: *How will you service from State 11?* for HPPO in Simple. At each stage, it gives an action and *beliefs over state-changes*. Green: added predicates; red: removed predicates. The header row is: Action ($I_d(s')$).

In the case of Simple, both agent pairs appear to specialise, each agent focusing on different desires: Green cooks, Blue serves; probably due to the difficulties of coordinated cooking in the cramped space. RAg scores extremely poorly but manages to fulfil some desires very infrequently (likely due to the small state-space). While for *Pair A* the agents are completely specialised, never doing the other task, for *Pair B*'s HAg policy learnt to serve infrequently and the two agents do both roles, though it is an unlikely occurrence.

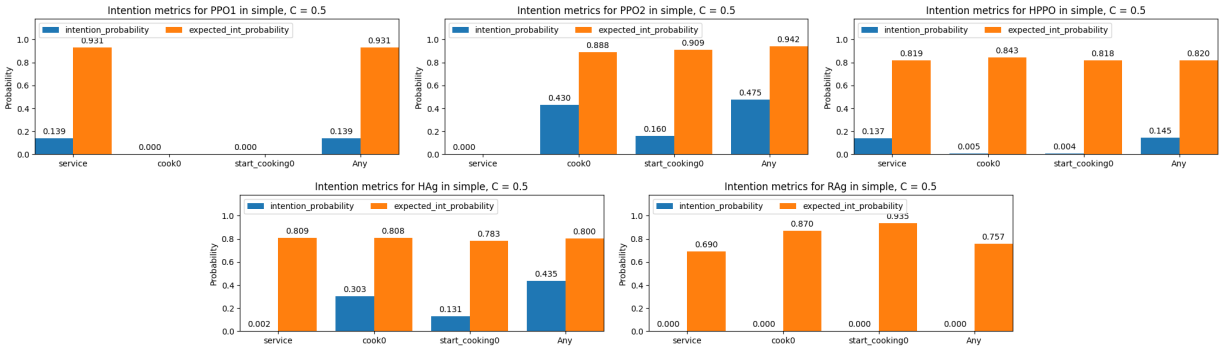
In the case of Random 0, *Green* agents are unable to fulfil any desire as they are trapped on the left aisle. The other two useful agents (PPO1, HPPO) show clear differences in interpretability of behaviour, and thus apparent rationality. When examining under the real-time explainability –or a good number of episodes– it becomes apparent that this is related to HPPO being frequently blocked by HAg: 80% of the time, HPPO is waiting for onions to place in the pot, and no intentional behaviour can exist without affordances making desires viable.

4.4 Example XAI Outputs

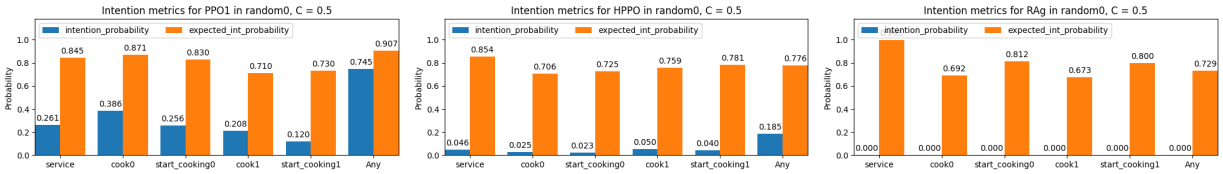
Finally, example explanations are produced via the algorithms in § 3.1.2. Let us select randomly a relatively common state (with attributed intentions) from the HPPO-Simple *PG*: State 11. Its predicates describe that it holds a dish, it is in front of the pot –which is done cooking– with the rest of predicates describing relative positions of the depots and other agent being southwest.

With a C of 50%, the answer to *what* it intends to do in $S = 11$ is *service* ($I_{service} = 0.505$). From this action, only Interact, Down and Up actions have been observed. The answer *Why* do Interact in $S = 11$ is that it furthers service as it increases $I_{service}$ by 0.08. For *How* to achieve *service* from $S = 11$, Table 3 shows the reply, showing the sequence of actions and believed transitions. In this case, further analysis of the *PG* unveils the reason for low (50.5%) intention: as the position of the other agent close to the plates means it could be grabbing one, and HAg often does this. Unable to pick soup and stuck with a dish, HAg stops cooking soup, resulting in suboptimal behaviour, and thus HPPO often (49.5% of the time) aborts serving in favour of letting the other agent do so.

Finally, a trajectory is studied in Figure 4. In this graph, the intentions of the agent through time ($I_d(s^t)$) predicts behaviour at a distance. Periods of no-intention are followed by abrupt increases, coinciding with HAg passing items over the counter.



(a) Intention metrics for the Simple Layout



(b) Intention metrics for the Random 0 Layout; PPO2 and HAg agents were intentionless and thus removed

Figure 2: Intention metrics for the 5 agents (in order, PPO1, PPO2, HPPO, HAg, and RAg) and $C = 0.5$.

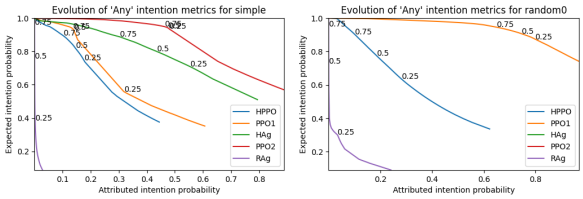


Figure 3: Attributed intention probability and expected intention probability over commitment thresholds in layouts Simple and Random 0. AUC represents explainability.

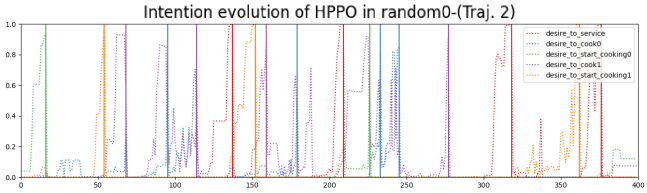


Figure 4: Intentions of HPPO in Random 0 example. Intention progression is marked with dotted lines, and desire completion with vertical solid lines. Each colour represents a desire.

5 DISCUSSION AND FUTURE WORK

Introducing intentions into *XAI* offers new vocabulary to do queries to an opaque model. We introduce metrics for hypothesising and testing behavioural hypotheses in the form of desires, offering insights into behaviour. The causal nature of intentions –and the state-changes they produce– allow to chain questions with concise answers, allowing for succinct explainability. In this paper, we focus on the formal part and intuitive usage of intentions, but also offer three possible *XAI* queries and algorithms for doing explainability from a telic, intention-based perspective, plus a real-time visualisation for observing intention progression through time.

We remark that the objective of this paper is to give tools to characterise the agent, opening mechanisms to identify rational versus irrational or random behaviour from the perspective of an explainee. Notably, the explainee need not be a human, not does *XAI* need to be about justifying behaviour, and the current limitations of the method can be exploited to feed new algorithms.

For example, as unseen transitions, or state regions with low number of observations decrease the *PG* interpretability –cannot reason about unseen transitions–, with access to the environment one can create a self-correcting policy that creates goals to explore unobserved parts of the state-action-space, creating an *interventional* world model $P(s'|do(a), s)$ [41] which qualitatively improves information over observation alone and would allow more counterfactual explanations. Similarly, an opposite direction can be explored by removing or discouraging actions that are not explainable from the intention perspective, pruning the policy and potentially increasing agent rationality.

Finally, there are two main limitations which this work does not address. Firstly, to fine-tune explainability for human-desirability, human studies may be needed; which have remained outside the scope of this work as we focused on other facets of intention usefulness. Secondly, currently we only examine desires of the *perform* or *achievement* type, though we remain optimistic about extending these to *maintenance* [58].

ACKNOWLEDGEMENTS

This work has been partially supported by the AI4EUROPE (Grant agreement ID: 101070000), SoBigData PPP (Grant agreement ID: 101079043) and V. Gimenez-Abalos' fellowship within the "Generación D" initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent atraction (C005/24-ED CV1). Funded by the European Union NextGenerationEU funds, through PRTR.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Silvia Tulli Aha, David W. (Ed.). 2024. *Explainable Agency in Artificial Intelligence: Research and Practice.* CRC Press, Boca Raton, FL, USA. https://doi.org/10.1201/9781003355281

[3] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (May 2018), 66–95. https://doi.org/10.1016/j.artint.2018.01.002

[4] Christian Arzate Cruz and Takeo Igarashi. 2020. A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference.* ACM, Eindhoven Netherlands, 1195–1209. https://doi.org/10.1145/3357236.3395525

[5] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2020. On the Utility of Learning about Humans for Human-AI Coordination. https://doi.org/10.48550/arXiv.1910.05789 arXiv:1910.05789 [cs, stat].

[6] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? https://doi.org/10.48550/ARXIV.2307.09009 Publisher: arXiv Version Number: 3.

[7] Giovanni Ciatto, Roberta Calegari, Andrea Omicini, Davide Calvaresi, and others. 2019. Towards XMAS: explainability through multi-agent systems. In *CEUR WORKSHOP PROCEEDINGS*, Vol. 2502. Sun SITE Central Europe, RWTH Aachen University, Rende, Italy, 40–53.

[8] Giovanni Ciatto, Michael I. Schumacher, Andrea Omicini, and Davide Calvaresi. 2020. Agent-Based Explanations in AI: Towards an Abstract Framework. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling (Eds.). Vol. 12175. Springer International Publishing, Cham, 3–20. https://doi.org/10.1007/978-3-030-51924-7_1 Series Title: Lecture Notes in Computer Science.

[9] Philip R Cohen and Hector J Levesque. 1990. Intention is choice with commitment. *Artificial intelligence* 42, 2-3 (1990), 213–261. Publisher: Elsevier.

[10] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence* 299 (Oct. 2021), 103525. https://doi.org/10.1016/j.artint.2021.103525 arXiv:2107.03178 [cs].

[11] Marc Domenech I Vila, Dmitry Gnatyshak, Adrian Tormos, Victor Gimenez-Abalos, and Sergio Alvarez-Napagao. 2024. Explaining the Behaviour of Reinforcement Learning Agents in a Multi-Agent Cooperative Environment Using Policy Graphs. *Electronics* 13, 3 (Jan. 2024), 573. https://doi.org/10.3390/electronics13030573

[12] Marc Domènech i Vila, Dmitry Gnatyshak, Adrián Tormos, and Sergio Alvarez-Napagao. 2022. Testing Reinforcement Learning Explainability Methods in a Multi-Agent Cooperative Environment. In *Frontiers in Artificial Intelligence and Applications*, Atia Cortés, Francisco Grimaldo, and Tommaso Flaminio (Eds.). IOS Press, Sitges, Catalonia, 355–364. https://doi.org/10.3233/FAIA220358

[13] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (Dec. 2019), 68–77. https://doi.org/10.1145/3359786

[14] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable Planning. https://doi.org/10.48550/arXiv.1709.10256 arXiv:1709.10256 [cs].

[15] Stan Franklin and Art Graesser. 1997. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent Agents III Agent Theories, Architectures, and Languages*, J. G. Carbonell, J. Siekmann, G. Goos, J. Hartmanis, J. Van Leeuwen, Jörg P. Müller, Michael J. Wooldridge, and Nicholas R. Jennings (Eds.). Vol. 1193. Springer Berlin Heidelberg, Berlin, Heidelberg, 21–35. https://doi.org/10.1007/BFb0013570 Series Title: Lecture Notes in Computer Science.

[16] Daniel Freeman, David Ha, and Luke Metz. 2019. Learning to predict without looking ahead: World models without forward prediction. *Advances in Neural Information Processing Systems* 32 (2019).

[17] Maor Gaon and Ronen Brafman. 2020. Reinforcement learning with non-markovian rewards. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 3980–3987. Issue: 04.

[18] Victor Gimenez-Abalos. 2024. Toward Explainable Agent Behaviour. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2740–2742.

[19] Victor Gimenez-Abalos, Luis Oliva-Felipe, Javier Vázquez-Salceda, Ulises Cortés, and Sergio Alvarez-Napagao. 2024. Why Interpreting Intent Is Key for Trustworthiness in the Age of Opaque Agents. https://doi.org/10.20944/preprints202402.1446.v1 Publisher: Preprints.

[20] H. P. Grice. 1975. Logic and Conversation. In *Speech Acts*, Peter Cole and Jerry L. Morgan (Eds.). BRILL, Leiden, The Netherlands, 41–58. https://doi.org/10.1163/9789004368811_003

[21] Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2023. Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. https://doi.org/10.48550/ARXIV.2302.10809 Publisher: arXiv Version Number: 3.

[22] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation* 16, 1 (Jan. 2024), 45–74. https://doi.org/10.1007/s12559-023-10179-8

[23] 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, Vienna Austria, 303–312. https://doi.org/10.1145/2909824.3020233

[24] Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (April 1944), 243. https://doi.org/10.2307/1416950

[25] Mark K. Ho, Rebecca Saxe, and Fiery Cushman. 2022. Planning with Theory of Mind. *Trends in Cognitive Sciences* 26, 11 (Nov. 2022), 959–971. https://doi.org/10.1016/j.tics.2022.08.003

[26] Monte Ransome Johnson. 2005. Teleology and Humans. In *Aristotle on Teleology* (1 ed.). Oxford University Press, Oxford, UK, 211–246. https://doi.org/10.1093/0199285306.003.0009

[27] Pat Langley. 2024. From Explainable to Justified Agency. In *Explainable Agency in Artificial Intelligence.* CRC Press, Boca Raton, FL, USA, 20.

[28] David Lewis. 1986. Causal explanation. In *Philosophical Papers Vol. Ii*, David Lewis (Ed.). Vol. 2. Oxford University Press, New York, NY, USA, 214–240.

[29] Zachary C. Lipton. 2017. The Mythos of Model Interpretability. http://arxiv.org/abs/1606.03490 arXiv:1606.03490 [cs, stat].

[30] Tongtong Liu, Joe McCalmon, Thai Le, Md Asifur Rahman, Dongwon Lee, and Sarra Alqahtani. 2023. A novel policy-graph approach with natural language and counterfactual abstractions for explaining reinforcement learning agents. *Autonomous Agents and Multi-Agent Systems* 37, 2 (Aug. 2023), 34. https://doi.org/10.1007/s10458-023-09615-8

[31] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2023. Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. https://doi.org/10.48550/arXiv.2310.19775 arXiv:2310.19775 [cs].

[32] Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. 2020. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction.* Springer, Dublin, Ireland, 1–16.

[33] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[34] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable Reinforcement Learning through a Causal Lens. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 03 (April 2020), 2493–2500. https://doi.org/10.1609/aaai.v34i03.5631

[35] Bertram F. Malle. 2022. Attribution theories: How people make sense of behavior. In *Theories in Social Psychology, Derek Chadee (Ed.)* (2nd edition ed.). John Wiley & Sons Ltd, Hoboken, NJ, US, 93–119.

[36] Bertram F. Malle and Joshua Knobe. 1997. The Folk Concept of Intentionality. *Journal of Experimental Social Psychology* 33, 2 (March 1997), 101–121. https://doi.org/10.1006/jesp.1996.1314

[37] Bertram F. Malle and Joshua Knobe. 1997. Which behaviors do people explain? A basic actor–observer asymmetry. *Journal of Personality and Social Psychology* 72, 2 (Feb. 1997), 288–304. https://doi.org/10.1037/0022-3514.72.2.288

[38] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A Survey of Explainable Reinforcement Learning. http://arxiv.org/abs/2202.08434 arXiv:2202.08434 [cs].

[39] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[40] Woosuk Park. 2022. How to Make AlphaGo's Children Explainable. *Philosophies* 7, 3 (June 2022), 55. https://doi.org/10.3390/philosophies7030055 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[41] Judea Pearl. 2000. *Causality: models, reasoning, and inference.* Cambridge University Press, Cambridge, U.K. ; New York.

[42] Jairo Perez-Osorio and Agnieszka Wykowska. 2020. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology* 33, 3 (April 2020), 369–395. https://doi.org/10.1080/09515089.2019.1688778

[43] Erika Puiutta and Eric MSP Veith. 2020. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction.* Springer, Dublin, Ireland, 77–95.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.48550/arXiv.1602.04938 arXiv:1602.04938 [cs, stat].

[45] Jan Robine, Tobias Uelwer, and Stefan Harmeling. 2023. Smaller World Models for Reinforcement Learning. *Neural Processing Letters* 55, 8 (Dec. 2023), 11397–11427. https://doi.org/10.1007/s11063-023-11381-3

[46] Bruno Rodrigues, Matthias Knorr, Ludwig Krippahl, and Ricardo Gonçalves. 2023. Towards Explaining Actions of Learning Agents. In *Proc. of Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (Eds.)*. London, UK, 9. https://alaworkshop2023.github.io/

[47] Bidipta Sarkar, Aditi Talati, Andy Shih, and Dorsa Sadigh. 2022. PantheonRL: A MARL Library for Dynamic Training Interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13221–13223. https://doi.org/10.1609/aaai.v36i11.21734 ISSN: 2374-3468, 2159-5399 Issue: 11 Journal Abbreviation: AAAI.

[48] Laura Sartori and Andreas Theodorou. 2022. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology* 24, 1 (Jan. 2022), 4. https://doi.org/10.1007/s10676-022-09624-3

[49] Kristin E. Schaefer, Edward R. Straub, Jessie Y.C. Chen, Joe Putney, and A.W. Evans. 2017. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research* 46 (Dec. 2017), 26–39. https://doi.org/10.1016/j.cogsys.2017.02.002

[50] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. https://doi.org/10.48550/arXiv.1707.06347 arXiv:1707.06347 [cs].

[51] John R. Searle. 1980. The Intentionality of Intention and Action. *Cognitive Science* 4, 1 (Jan. 1980), 47–70. https://doi.org/10.1207/s15516709cog0401_3

[52] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, 618–626. https://doi.org/10.1109/ICCV.2017.74

[53] Ben R. Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P. Ginsburg. 1993. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology* 23, 3 (May 1993), 219–238. https://doi.org/10.1002/ejsp.2420230302

[54] James Somers. 2018. How the artificial-intelligence program AlphaZero mastered its games. *The New Yorker* 3 (2018).

[55] Aaquib Tabrez and Bradley Hayes. 2019. Improving Human-Robot Interaction Through Explainable Reinforcement Learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, Korea (South), 751–753. https://doi.org/10.1109/HRI.2019.8673198

[56] Adrian Tormos, Victor Gimenez-Abalos, Javier Vázquez-Salceda, and Sergio Alvarez-Napagao. 2024. pgeon applied to Overcooked-AI to explain agents' behaviour. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2821–2823. event-place: Auckland, New Zealand.

[57] Adrián Tormos, Víctor Giménez Ábalos, Marc Domènech Vila, Dmitry Gnatyshak, Sergio Álvarez Napagao, and Javier Vázquez Salceda. 2023. Explainable agents adapt to human behaviour. In *Proceedings of the First International Workshop on Citizen-Centric Multi-Agent Systems (CMAS'23)*. 42–48. https://upcommons.upc.edu/handle/2117/390757

[58] M. Birna van Riemsdijk, Mehdi Dastani, and Michael Winikoff. 2008. Goals in agent systems: a unifying framework. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2 (AAMAS '08)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 713–720. event-place: Estoril, Portugal.

[59] Pulkit Verma, Shashank Rao Marpally, and Siddharth Srivastava. 2022. Discovering User-Interpretable Capabilities of Black-Box Planning Agents. In *Proceedings of the Nineteenth International Conference on Principles of Knowledge Representation and Reasoning*. International Joint Conferences on Artificial Intelligence Organization, Haifa, Israel, 362–372. https://doi.org/10.24963/kr.2022/36

[60] Michael Winikoff, Virginia Dignum, and Frank Dignum. 2018. Why Bad Coffee? Explaining Agent Plans with Valuings. In *Developments in Language Theory*, Mizuho Hoshi and Shinnosuke Seki (Eds.). Vol. 11088. Springer International Publishing, Cham, 521–534. https://doi.org/10.1007/978-3-319-99229-7_47 Series Title: Lecture Notes in Computer Science.

[61] Michael Winikoff and Galina Sidorenko. 2023. Evaluating a Mechanism for Explaining BDI Agent Behaviour. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2283–2285. event-place: London, United Kingdom.

[62] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What does the robot think? Transparency as a fundamental design requirement for intelligent systems. In *IJCAI 2016 Ethics for AI Workshop*.

[63] Georg Henrik von Wright. 2004. *Explanation and Understanding*. Cornell University Press, Ithaca, NY, USA. Google-Books-ID: 33wCi2bg5x0C.

[64] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* 5, 5 (Oct. 2021), 726–742. https://doi.org/10.1109/TETCI.2021.3100641 arXiv:2012.14261 [cs].

[65] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5 (Jan. 2021), 593. https://doi.org/10.3390/electronics10050593 Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.