Towards Explainable Agent Behaviour

Doctoral Consortium

Victor Gimenez-Abalos
Barcelona Supercomputing Center / Universitat Politecnica de Catalunya
Barcelona, Spain
victor.gimenez@bsc.es

ABSTRACT

Agents are a special kind of AI-based software in that they interact in complex environments and have increased potential for emergent behaviour, even in isolation. Explaining such behaviour is key to deploying trustworthy AI, but the increasing complexity and opaqueness of agents makes this hard. Beyond narrow-task and instant-based goals, agents may exhibit durative behaviour and be required to have planning or deliberative capabilities, or even to reason over other's behaviours. This precludes machine learning explainability -i.e. explanations over single predictions or actions- from giving complete and useful explanations. There is a need for extending explainability tools. We split the capabilities of agents into several levels, each more abstract, and produce explanations by climbing these levels: from actions, tellic (ends), deliberation, and more. The first two have been solved through frequentist models (Policy-Graphs), and the third is work in progress. We intend to extend this work by adding components for explaining epistemology, agent-agent interaction, norms and values.

KEYWORDS

explainability; theory of mind; intention inference; planning; desires; deliberation; multi-agent systems; values alignment

ACM Reference Format:
Victor Gimenez-Abalos. 2024. Towards Explainable Agent Behaviour: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 EXTENDED ABSTRACT

Among the tasks within the purview of Artificial Intelligence (AI), the issue of solving problems without giving explicit knowledge on *how* to solve them is very 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 [8, 15]. This is where the field of *Explainable Artificial Intelligence* (*XAI*) shines through.

A model explanation can be understood as a communication between an explainer algorithm (be it intrinsic or extrinsic to the explained component) and a receiver or explainee (which can be a human or a component for a downstream task). The content of an explanation should adhere to several principles to describe the relevant context or the causes surrounding some facts [7, 11, 14], and in the context of AI it is often related to its intermediary or final decisions. In order for the communication to be effective, four main maxima should be followed [5]: the message should be comprehensible to the receiver (interpretability), it ought to contain truthful information from the perspective of the explainer (reliability), its length should be just enough to be informative, and its content should be relevant to the context.

In the context of today's *XAI*, interpretability is often the main metric considered when evaluating the 'explainability' of a model or method. However, the fact that interpreted outputs of an explainability algorithm (no matter how clear) may be confabulation of an explainability method precludes these explanations from being trustworthy. Some metrics have already been introduced to evaluate reliability for vision [2].

However, when exporting traditional machine learning explanations to reinforcement learning agents, what should the content and extent of explanations be? Common methods may give explanations regarding the choice of single actions as a relevance of the percepts of the model or internal variables. However, these explanations often lag behind in the necessities of the explainee, as courses of action cannot be understood from state relevance maps [3, 13]. Instead, the burden of inferring those falls on the explainee, possibly arriving at wrong conclusions based on anthropomorphising an agent that may learn in very different ways from our own.

Interpretability and reliability can be seen as two separate optimisation objectives, which tend to be in conflict. For example, the most reliable explanation of an opaque model would be a complete specification of its code, whilst the more interpretable would be its most abstract, oversimplified and potentially misleading description of its behaviour. That being said, the validity of each explanation depends on the context and receiver: for example, a student learning about architectures may find the code specification more desirable than abstract explanations. As a community we need explainability techniques that can optimise these criteria, while allowing the explainee to choose how they are prioritised. Furthermore, the heterogeneity of agent architectures, as well as the opaqueness of some models, means that explainability techniques often sacrifice generality to deal with particular agents. We propose to search for algorithms that work agnostic to agent architecture, and build their own representation based on external observation alone like a human would, only assuming we are able to observe (part of) their state or percepts and their actions.

In our works so far, we have focused on the extension of a simple, frequentist technique we call *Policy Graph* (*PG*). Introduced by Hayes et. al. [6], these models were used for producing natural language explanations over agent behaviour, constrained to explaining behaviour from the scope of single actions but extended to reply

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

questions of behaviour as a whole (such as 'in which conditions do you perform this action'). The model is built as a probabilistic graphical model, where both the policy of the agent ($P(a|s)$) and how it affects the environment ($P(s'|a,s)$) are estimated by external observations, making it a post-hoc explainability technique.

These explainer models were converted into a simulacra of original, capable of taking actions in the environment [4]. This transformation creates two main contributions. On the one hand, the simulated agent can be deployed as the final agent while being transparent making it explainable (on actions) by design. This agent, however, may under-perform compared to the original, especially if its world representation or amount of observations were insufficient to capture the original policy. On the other hand, we pose that the gap in performance between the original agent and its simulacra offers a good metric of the reliability of explanations proposed by a *PG*: if both agents obtain the same level of performance, their behaviour critical to the task is similar enough and thus the explanations the *PG* offers with regards to it must capture the same processes the original agent does.

In follow-up work [12], we focus on collaborative agents, focusing on extracting explanations and metrics when one of the agents is the simulacra of a human, and seeing how, in the cases where collaboration is required, reliability of explanations falls when the world representation lacks information required to cooperate with a human. In addition to all of these, we have created and presented an open repository for the implementation of our techniques [1], which we plan to extend with our future findings.

One of the main drawbacks of the model is that, currently, the *PG* requires a discretisation of states to be applied, since original state-spaces are too large to use frequentist methods to estimate transition probabilities from a reasonable amount of samples. In parallel to our ongoing work, we are attempting to shift away from discretisation of the state-space by adding formalisms that transform the stored probability distribution over discretised states ($P(s', a|s)$) into something more sophisticated, not considering single states or actions but salient features of states or 'skills' ($P(f(s'), g(skill)|f(s))$, deriving inspiration from intrinsically motivated learning and off-line learning work [10]. Machine-learning methods could serve to bridge the gap, but can compromise the reliability of frequentist methods, as with the current approach we know why the model believes $P(s', a|s)$ (it was empirically observed). The hypothesis is that this may eventually be solved by new epistemic engines that can learn characteristics of the effects of actions in the environment, and can reply to questions of why they believe so.

Our ongoing work, is on climbing the ladder of explanations to provide more abstract reasoning over actions [3]. In an in-progress journal paper, we have extended *PG*s beyond giving explanations of single actions as a function of the state. We propose tellic explanations by introducing concepts of BDI architectures (desires and intentions), that allow *PG*s to answer questions such as *With what purpose have you performed this action?* to which we can reply: *To fulfil this objective*, or *How do you plan to achieve this objective?*: *By performing this course of actions which should change the states in this manner*. In addition to algorithms for answering these questions, we propose a set of metrics that evaluate both interpretability of behaviour overall and reliability of the explanations provided, and a threshold for tuning the trade-off between the two metrics. Furthermore, the explanation algorithms can be composed, allowing explanations of increasing level of detail and length based on making further questions about the replies provided.

Based on work performed in the European Project HumaneAI, we are building a ladder of explanations, in which each level contains the causes of the behaviours produced above it. Starting at action scope, actions are generated due to having a policy and being in a state. From there, one can question further by asking why does the agent have that policy. This scales the question up to the planning scope in which the explanation ought to include the intention of the action. Understood from the point of view of folk-conceptual theory of explanations, explanations involving intention [9] include both the finality of the action, and beliefs about how that finality can be brought about by the action. In our current stage, with tellic *PG*s, we believe these explanations are already satisfactorily extracted. For example, in the game overcooked, we are able to ask an agent *What do you intend to do?*, to which it can reply *I intend to deliver soup*; *How do you plan to deliver soup?*: *I will move right, putting me in this state (which is different from the previous one in these qualities), after which I will go down putting me in this other state, and then I will interact, causing me to deliver soup.* We have not yet created simulacra of agents that take these additions into consideration, using it as a diagnostic tool so far. Another question we intend to solve when doing this would be *When is an intention manifested?*, to tie intention to affordances and world states back again. For example, the intention to serve soup would start when I have access to the plates and a pot that has soup.

In turn, explanations at behaviour level ought to be expanded by allowing to query about the finality of a finality (or *Causal History of Reasons* [9]). This would be tied to both learning the deliberation process of another agent, and which epistemic mechanisms have brought about the beliefs on how the world changes via the agent's interactions. Our target in this stage would be being able to answer questions such as: *Why did you prioritise delivering soup over putting an onion in the pot?*, or *Why did you believe that going up would put you in a state in which the pot is on your left?*. This part of the project is our current research focus: trying to extract the capabilities of agents which do this (implicitly or explicitly) from observations, and storing this information in a way that an algorithm may answer these questions. We are also considering adding the requirement of being able to mimic these capabilities much like with previous agent, which would entail creating a transparent architecture that uses information on epistemic, deliberation, planning, and action capabilities to create behaviour.

Further down the line, we consider extending the framework to consider agents with *Theory of Mind* (*ToM*) 1 and 3, involving questions about how the world changes in the presence of another agent with goals, or why some desires were discarded (by the presence of norms). Another research line we are considering is making opaque agents (RL) that have an internal *PG* model of another agent as part of their input to improve their *ToM*1 capabilities.

As higher levels are explored, we consider creating a cognitive architecture based on the lessons learnt on how to produce explanations on behaviour. For example, can insights from how to answer explanations on the origin of a learnt world model be used to build an 'experimenter' agent that sets itself goals about acquiring knowledge?

REFERENCES

[1] Sergio Alvarez-Napagao, Adrián Tormos, Victor Abalos, and Dmitry Gnatyshak. 2023. Policy graphs in action: explaining single- and multi-agent behaviour using predicates. In *XAI in Action: Past, Present, and Future Applications*. https://openreview.net/forum?id=QPqL9xsYOf

[2] Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Giménez-Ábalos. 2022. Focus! Rating XAI Methods and Finding Biases. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–8. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882821

[3] 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].

[4] 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. https://doi.org/10.3233/FAIA220358

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

[6] 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

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

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

[9] 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

[10] Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Alexandre Lacoste, and Sai Rajeswar. 2022. Choreographer: Learning and Adapting Skills in Imagination. https://openreview.net/forum?id=PhkWyijGi5b

[11] 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

[12] Adrián Tormos Llorente, 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

[13] 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.

[14] Georg Henrik von Wright. 2004. *Explanation and Understanding*. Cornell University Press. Google-Books-ID: 33wCi2bg5x0C.

[15] 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].