Explaining Agent Intentions

Doctoral Consortium

Sara Montese
Barcelona Supercomputing Center / Universitat Politecnica de Catalunya
Barcelona, Spain
sara.montese@bsc.es

ABSTRACT

Understanding *why* an agent acts in a certain way based purely on its observed behaviour is not an easy hill to climb, especially when explanations aspire to be reliable and interpretable. Despite the growing body of research on Explainable Agency (XAg), existing approaches to agent explainability often remain disconnected from causal and teleological theories of explanation or rely on the strong assumption that agents pursue a single goal or a set of independent goals.

My thesis aims to develop methods to explain the reasoning behind *when*, *how*, and *why* agents pursue and prioritise multiple intentions, and how these intentions relate to one another.

We first discuss what we mean by explainability, and we introduce a first step to extract the agent's priorities from observed behaviour. We then outline future directions for using this information to support policy improvement and to infer inter-intention relationships, and open a discussion on the boundaries of agent explainability.

KEYWORDS

Explainable AI; Agent Explainability; Intentions; Causality

ACM Reference Format:
Sara Montese. 2026. Explaining Agent Intentions: Doctoral Consortium. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/XMCQ3388

1 EXTENDED ABSTRACT

Consider an *Autonomous Vehicle* (AV) navigating in an urban area. While travelling along a road and preparing to shift to the middle-left lane, a second vehicle suddenly attempts to merge from the left (Figure 1). In response, the AV applies emergency braking and halts. Alternatively, it could have decelerated smoothly, maintained its velocity under the expectation that the merging vehicle would yield, or adopted another evasive strategy. Each of these courses of action represents plausible alternatives. Why, then, did the AV behave the way it did?

As autonomous agents increasingly rely on complex and opaque decision-making, explaining their behaviour sets out a technical challenge and a legal requirement [7]. The field of *Explainable Artificial Intelligence* (*XAI*) aims to address this challenge by providing explanations that support the justification and accountability of the system, assist its control and improvement, and encourage the discovery of new knowledge [1].

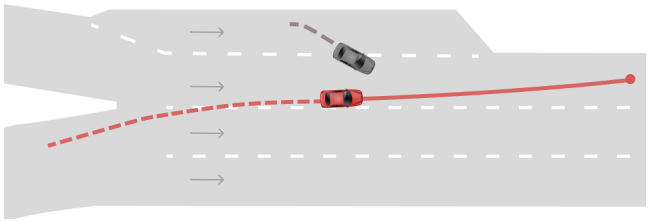

Figure 1: The AV (in red) is transitioning to the middle-left lane of the road. During the manoeuvre, another vehicle recklessly enters into the AV's trajectory.

Two fundamental properties can be identified for any explainability system:

- **Causality**: An explanation of an observed behaviour is, indeed, the selection and presentation of *causal factors* behind that behaviour, in response to a *why*-question [15, 20, 25].
- **Conversationality**: Since the act of explaining develops as a *conversation* of questions and answers between an explainer (*e.g.* the *XAI* model) and the explainee (*e.g.* a human user), it must ideally conform to Grice's conversational maxims, and therefore be concise, reliable, relevant, and understandable by the explainee [9, 13].

Most traditional *XAI* methods for supervised learning models explain system decisions by highlighting input features, weights, or pixels deemed relevant during the decision-making [2, 21]. However, these approaches fail to meet users' needs and expectations [17], not least when applied to *Explainable Agency* (*XAg*), the subfield of *XAI* dedicated to explaining agent behaviour. This explanatory gap calls for a refocus on the *human* explainee.

Humans explain others' actions by attributing mental states such as beliefs, desires, and intentions, and when the behaviour is perceived as intentional, people most commonly cite the actor's *intentions* as the cause of the behaviour [4, 18]. Given people's tendency to anthropomorphise artificial agents [12], the explanatory framework for human actions can be adapted for artificial agents, positioning **Teleology** as a third pillar in *XAg* [8, 10].

Building on these properties, the *ladder of intentions* [6] has been proposed to unify the explainability questions of agent behaviour. This framework, built upon *Belief-Desire-Intention* (BDI) concepts, decomposes the agent architecture into components at different levels of abstraction, from the highest level (the designer's intentions) to the lowest level (the executed action). Each level represents the

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/XMCQ3388

causes of the level below, and explanations can be extracted by posing questions at each level. For example, starting from observed actions (lowest level), *XAg* questions that arise are why an action was chosen, and why the agent believed it would fulfil its policy or plan of action according to its beliefs.

At a higher level, *XAg* queries concern why the agent was following a given policy and why it believed the policy fulfilled its desires through the generated intentions. At a third level, explanations concern the agent's prioritisation among different desires, asking why an intention was formed and why the agent believed it satisfied its priorities.

Within the literature in *XAg* [19, 24], action-level explainability queries have largely been addressed through importance-based approaches, whereas fewer methods offer explanations apt to policy-level questions. In previous work, a *post-hoc* explainability method, *Intention-aware Policy Graph*s (*IPG*s) [5] was introduced as an intuitive approach to teleological explanations, *i.e.* explanations in terms of an agent's purpose(s). An *IPG* is a probabilistic graphical model constructed in a frequentist manner from observations of an agent interacting within an environment. Each node of the graph corresponds to a discretised state s, and each edge represents the probability $P(s', a|s) = P(a|s) * P(s'|s, a)$ of transitioning from state s to state s' upon executing action a.

A desire d hypothesised to guide the agent's behaviour is formalised as the execution of a desirable action a_d within a desirable state region S_d. The agent's intention $I_d(s)$ to fulfil a desire d from a state s is defined as the probability of reaching a state in S_d and executing a_d, starting from s. This intention value is computed as the sum of probabilities of all possible sequences of states and actions (paths) starting from s and culminating in (s', a_d), where $s' \in S_d$.

Based on this approach, we applied *IPG*s to provide teleological explanations for the behaviour of an AV in urban environments [22, 23], using data from real driving scenes. After defining the vehicle's possible desires (*e.g.* to stop at a stop sign or to allow a pedestrian to cross), the method allows intention attribution and the extraction of purpose-oriented explanations for global driving behaviour and local driving decisions, as well as identifying anomalous or undesirable conduct (*e.g.* ignoring stop signs). Consider the scenario in Figure 1. Our method allows us to answer the question *"Why did the AV stop?"* by retrieving the intentions attributed to the vehicle at that state and identifying, with a quantifiable degree of reliability, the motivation behind its behaviour. For instance, the explanation might indicate that *"The AV stopped because it had a high intention to avoid collisions, and the action to stop brought about that intention".*

In our in-progress work, we aim to reach the third rung of the ladder framework to explain why an agent selects and brings about a particular intention based on the prioritisation of its desires, and why those priorities influence the intention-selection process.

To illustrate the value of such explanations, consider the scenario in Figure 1. The AV initially intends to execute a lane change to the left (I_1), while also maintaining safety (I_2). When a second vehicle unexpectedly enters the lane, the AV brakes abruptly. If an agent is asked to explain its behaviour, the AV response would be: *"I initially intended to change lane and maintain safety, as I believed these intentions could be pursued concurrently. Upon the detection of a nearby vehicle, I prioritised safety over lane-changing; therefore, I stopped."*

State-of-the-art methods that elicit purpose-oriented explanations (*e.g.* [3, 5, 11]) rely on the assumption that an agent pursues a single goal, or the most probable one(s) among a set of goals. However, an agent may deliberate to fulfil multiple intentions concurrently within a single sequence of actions, or to prioritise one intention over another, depending on its global and contextual priorities, and on the relationships between its desires. To the best of our knowledge, no method in the *XAg* literature addresses explainability queries about multi-intention reasoning.

We begin by improving the current formulation of intention in *IPG*s to distinguish between situations in which the agent is acting or not optimally towards a desire. An *intention* of a desire I_d is defined as the probability of successfully fulfilling the desire, influenced by two factors: the agent's motivation towards the desire, and its belief in the possibility of achieving it. This definition entangles the desirability and the feasibility of reaching the desirable situation, since the increase in intention from one to another considers both the agent's choice ($P(a|s)$) and the environment dynamics ($P(s'|s, a)$), limiting the explainee's ability to discern whether a high intention value arises from high desirability or high feasibility (or both).

In ongoing work, we address this conflation with the concept of *potential intention*. A potential intention of a desire PI_d is the maximum intention value I_d that the agent can attain if it were able to choose all actions until the desire is achieved.

Identifying which intention(s) an agent is actively pursuing supports the extraction of general and contextual priorities emerging from action choices. In future work, this information will be used to construct a graphical representation of inter-intention relationships, including independence (pursuing I_i does not influence I_j), conflict (pursuing I_i hinders I_j), and facilitation (pursuing I_i helps achieve I_j) [14]. Furthermore, future work aims at using potential intention values to identify suboptimal actions with respect to the agent's desires, and therefore to intervene on the graphical model to improve the agent's behaviour.

As we climb the levels of the ladder framework, the increasing integration of *Large Language Model*s (LLMs) into agent architectures raises questions about the boundaries of agent explainability. LLMs are trained to optimise autoregressive next-token prediction, rather than learning the meaning of language [16], which limits the reliability of causal threads between the model's inputs and its outputs in order to produce teleological explanations. This motivates new questions we aim to explore. How does the inclusion of LLM-components in the agent architecture affect the reliability of the explanations? Is it always possible to provide reliable explanations for the behaviour of an artificial agent?

ACKNOWLEDGMENTS

Sara Montese is supported by the fellowship within the "Generación D" initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (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). 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. [n.d.]. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. 99 ([n. d.]), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[3] Abeer Alshehri, Amal Abdulrahman, Hajar Alamri, Tim Miller, and Mor Vered. 2025. Towards Explainable Goal Recognition Using Weight of Evidence (WoE): A Human-Centered Approach. 82 (2025), 2535–2594. https://doi.org/10.1613/jair.1.17173

[4] Daniel C. Dennett. 1989. *The Intentional Stance.* MIT Press.

[5] 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 *Proceedings of the 24rd International Conference on Autonomous Agents and Multiagent Systems* (Richland, SC) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems. event-place: Detroit, United States of America.

[6] Victor Gimenez-Abalos, Adrian Tormos, Filip Edström, Sergio Alvarez-Napagao, Javier Vázquez-Salceda, Mattias Brännström, and John Lindqvist. 2026. Ladder of Intentions: Unifying Agent Architectures for Explainability and Transferability. In *Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems* (Cham), Davide Calvaresi, Amro Najjar, Andrea Omicini, Reyhan Aydogan, Rachele Carli, Giovanni Ciatto, Simona Tiribelli, and Kary Främling (Eds.). Springer Nature Switzerland, 127–146. https://doi.org/10.1007/978-3-032-01399-6_8

[7] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[8] M. D. Graaf and B. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *Proc. of the AAAI Fall Symposia 2017.* https://www.semanticscholar.org/paper/How-People-Explain-Action-(and-Autonomous-Systems-Graaf-Malle/22da5f6f70be46c8fbf233c51c9571f5985b69ab

[9] 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

[10] Balint Gyevnar, Stephanie Droop, Tadeg Quillien, Shay B. Cohen, Neil R. Bramley, Christopher G. Lucas, and Stefano V. Albrecht. 2025. People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior: Insights from Cognitive Science for Explainable AI. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 86, 18 pages. https://doi.org/10.1145/3706598.3713509

[11] Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2024. Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Richland, SC) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, 771–779. https://dl.acm.org/doi/10.5555/3635637.3662930

[12] F. Heider and M. Simmel. 1944. An experimental study of apparent behavior. 57 (1944), 243–259. https://doi.org/10.2307/1416950 Place: US Publisher: Univ of Illinois Press.

[13] Denis Hilton. 1990. Conversational Processes and Causal Explanation. 107 (1990), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

[14] Franki YH Kung and Abigail A Scholer. 2020. The pursuit of multiple goals. *Social and Personality Psychology Compass* 14, 1 (2020), e12509.

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

[16] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. [n.d.]. Dissociating language and thought in large language models. 28, 6 ([n. d.]), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

[17] Alessio Malizia and Fabio Paternò. 2023. Why Is the Current XAI Not Meeting the Expectations? 66, 12 (2023), 20–23. https://doi.org/10.1145/3588313

[18] Bertram F. Malle. 2004. *How the mind explains behavior: folk explanations, meaning, and social interaction.* MIT Press.

[19] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A Survey of Explainable Reinforcement Learning. https://doi.org/10.48550/arXiv.2202.08434 arXiv:2202.08434 [cs]

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

[21] Azza Mohamed, Khaled Abdelqader, and Khaled Shaalan. 2025. Explainable Artificial Intelligence: A Systematic Review of Progress and Challenges. (2025), 200595. https://doi.org/10.1016/j.iswa.2025.200595

[22] Sara Montese, Victor Gimenez-Abalos, Atia Cortés, and Ulises Cortés. 2025. Intention-aware Policy Graphs for Explainable Autonomous Driving. In *2025 IEEE Intelligent Vehicles Symposium (IV)* (Cluj-Napoca, Romania). 1928–1934. https://doi.org/10.1109/IV64158.2025.11097511

[23] Sara Montese, Victor Gimenez-Abalos, Atia Cortés, Ulises Cortés, and Sergio Alvarez-Napagao. 2026. Explaining Autonomous Vehicles with Intention-Aware Policy Graphs. In *Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems* (Cham), Davide Calvaresi, Amro Najjar, Andrea Omicini, Reyhan Aydogan, Rachele Carli, Giovanni Ciatto, Simona Tiribelli, and Kary Främling (Eds.). Springer Nature Switzerland, 40–57. https://doi.org/10.1007/978-3-032-01399-6_3

[24] Léo Saulières. 2025. A Survey of Explainable Reinforcement Learning: Targets, Methods and Needs. https://doi.org/10.48550/arXiv.2507.12599 arXiv:2507.12599 [cs]

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