Preprints.org

Article

Not peer-reviewed version

Why Interpreting Intent Is Key for Trustworthiness in the Age of Opaque Agents

Victor Gimenez-Abalos , Luis Oliva-Felipe , Javier Vázquez-Salceda , Ulises Cortés , Sergio Alvarez-Napagao
*

Posted Date: 26 February 2024

doi: 10.20944/preprints202402.1446.v1

Keywords: explainability, intention inference, values alignment, multi-agent systems

Article

Why Interpreting Intent Is Key for Trustworthiness in the Age of Opaque Agents

Victor Gimenez-Abalos [1], Luis Oliva-Felipe [2], Javier Vázquez-Salceda [2], Ulises Cortés [1,2] and Sergio Alvarez-Napagao [1,2,*]

[1] Barcelona Supercomputing Center, Barcelona, Spain; victor.gimenez@bsc.es (V.G.-A.); ia@cs.upc.edu (U.C.)
[2] Universitat Politècnica de Catalunya, Barcelona, Spain; luis.javier.oliva@upc.edu (L.O.-F.); jvazquez@cs.upc.edu (J.V.-S.)
* Correspondence: sergio.alvarez@bsc.es

Abstract: This paper addresses the critical issue of trust in Artificial Intelligence systems, especially when users might find it challenging to comprehend the internal decision-making processes of such systems. A relevant topic of research in this respect is Theory of Mind, which involves attempting to understand these systems as if they possessed beliefs, desires, and intentions. We focus on the latter: intentions, and we examine how producing explanations based on them can improve *understandability* while allowing for a better interpretation of how they align with human values. We also review some existing methods for identifying intentions in AI systems and we conclude with a discussion on possible future directions in this line of research.

Keywords: explainability; intention inference; values alignment; multi-agent systems

1. Introduction

Artificial Intelligence (AI) applications are widely used in many areas, such as health, finance, logistics, robotics, and transport. Understanding how AI-based systems make decisions is important [1], especially when they play a significant role in different sectors of society and can affect individual lives or the environment. However, the inner workings of these systems are often not easy to understand because they are complex and use complicated, or *opaque*, algorithms.

While it is relevant to acknowledge that AI-based systems do not possess beliefs, desires, or intentions because they do not have consciousness or subjective experiences, *theories of mind* upon them are often used. Through the Theory of Mind (see for example [2]), people attribute specific properties to the mental states of the other (human or AI) agents: beliefs, desires, intentions, emotions, and values. In many cases [3], this is used to help humans understand and predict better the actions of AI-based systems.

Much research has been done on *how* it is possible to attribute beliefs and desires [4,5] in AI-based systems. However, we argue that the attribution of intentions, esp. when there is no direct access to mental states, has yet to be comprehensively explored. In this paper, we discuss the importance of thinking and talking about artificial agents in terms of *intent* and how this can provide better mechanisms for explaining, understanding, and improving AI systems. In other words, we argue that *intentions* are essential to ensure and trust that such systems' actions are ethical and responsible.

1.1. Structure of this paper

In Section 2, we discuss the need for trustworthy agents in socio-technical systems and link this need to the Theory of Mind. In Section 3, we propose focusing on intentions for generating explanations that improve trust in opaque agents, and in Section 5, we analyse some attempts to produce such intentions. Finally, in Section 6, we summarise the paper and discuss future work.

2. The imperative for trustworthy agents: applying Theory of Mind

Trust in technology, particularly AI-based systems, is essential for effective adoption across various societal sectors. This trust is critical for those cases where users should be able to reliably predict and comprehend the behaviours and outputs of AI systems, *e.g.* where deployed in influential and impactful areas, such as in healthcare or autonomous transportation. Trust should not just be a product of a model's reliable and accurate performance but should also be influenced by applying ethical and responsibility guidelines and guardrails.

For achieving trustworthiness, one crucial need is for systems that are transparent and understandable by humans. This entails, at the very least, the necessity for having explainable AI-based systems to let users understand why a particular decision or action was taken, enabling stable expectations, confidence, and trust in the technology. This is in fact the focus of the broad topic of Explainable AI (XAI).

If we focus on *agents* A, that is, on *actors* that are situated in an environment, can sense it, and act upon it, several sources of data can be analysed for producing explanations about the actions of an agent. For example, if there is direct access to the agent's world model (or, in reinforcement learning, an understandable policy function or the reward function that produced the policy), creating an algorithm that explains the actions taken concerning the world model should be a simple task. However, that is not always the case, as in many cases, we might be dealing with opaque agents in the sense that they have a (private) policy that cannot be accessed or a policy that is too complex to be trivially understandable, such as in the case of deep neural network reinforcement learning. In other cases, we might be dealing with an environment that is so complex that the observations have to be heavily processed to be used for producing explanations.

Another important topic to cover when dealing with explanations is that not all explanations are desirable[1] for humans. When studying agents' behaviour, we argue that it is valuable for humans to understand the specific reasons behind an action taken by an Agent within the context of its medium- or long-term behaviour. Often, actions are part of a sequence or influenced by a combination of actions and observations, as well as external factors, rather than being isolated occurrences. In short, we propose that explainability should prioritise communicating reasons over behaviour, and substantiating explanations with not only desires and beliefs, but also intentions [6], rather than just enunciating how the agent is operating. This requires awareness of an agent's explicit mental states, desires, and goals.

Here, we propose to design mechanisms that can produce explanations that involve these mental states, even when they are not directly accessible. These states should include the specific goals the agent tried to achieve and originated the action. Even if the internal representation of the world (be it embedded in the code or the policy) that governs the agent is available, it is expected that it is not understandable by a human.

Our objective is to be able to extract knowledge from the observation of agents so that we can communicate their ongoing goals and how agents use actions to see to it that those goals are achieved to observers and stakeholders of the system. If we cannot assume that the mental states are accessible or understandable, and if we want to improve the trustworthiness of opaque agents, then we forcefully need to assume that we must find a way to infer them *via* mere observation of the agent in the environment.

In summary: understanding the decisions of opaque agents, especially when situated in complex environments, brings upon critical challenges: inaccessibility of internal mechanisms, the potential complexity of the decision-making process, the potential dynamic nature of the environment, the need for consideration of ethical guidelines, and being able to produce explanations that are desirable by humans.

[1] Another important topic to cover when dealing with explanations is that not all are understandable by humans.

Developing a *theory of mind* for agents is a possible way to tackle these issues. By interpreting an agent's behaviour in terms of beliefs and desires, humans might have more natural mechanisms to lower the complexity of the decision-making process or the relationship between the agent and the environment. Additionally, by referring to concepts such as beliefs and desires, it should become easier to check against ethical constraints or guidelines. This is what certain works in the literature of neuroscience [7–12] in the study of applying the *theory of mind* to humans have been focusing on for a considerable amount of time with interesting empirical results.

Similar approaches have been followed in AI research for agents, the most popular formalisation being the Belief-Desire-Intention agent architecture [13]. Agents built using this architecture produce a behaviour that is immediately transparent based on a set of internal rules, and it is straightforward to produce explanations that are desirable for humans due to these rules being implemented in terms of beliefs, desires, goals, intentions, plans, and actions all connected by an underlying logic [14].

3. The role of intention in explainability

Explainability is a communication process between an emitter and a receiver (also called explainee). Therefore, explanations are elements of language in a specific codification. The reliability of an explanation depends on the emitter, while the interpretability relies on the receiver.

Therefore, the codification they share through interaction must be robust and efficient. According to Grice's maxims of communication [15], explanations should be as minimal as possible. There needs to be a mechanism that is able to break down a question into further questions, but only if and when clarification is needed. For example: *Why did you, agent A, do X?* should be replied just with what the agent A believes the explainee wants to know, *e.g. Why did I move up? Because I wanted a sponge*, reply with more information if prompted, *e.g. Why did you move up for the sponge?*, replied with *I believe that the sponges are up*, *Why believe that?*, *Because I see the sponge in the upper part of my visual input*, etc.

Intentions can help structure or abstract the explanations in a way that they are more succinct, summarising behaviour and focusing on the content the (human) explainee would rather receive without restricting the explanation to a single level of abstraction. For example: *Why did I move up? Because I wanted to clean something*. After all, as stated in [16], to explain rational behaviour or the illusion of it, we need more than just beliefs and desires: we also need intentions. We propose to see intentions, whether explicit or implicit in the agent's decision-making process, as a desirable feature to build explanations upon and a final objective of any explainability pipeline for agents.

Understanding intention in AI-based systems usually requires a precise analysis of *how* these systems make decisions and the possible purposes or objectives behind these decisions. In this context, *intention* refers to an active and persistent goal that an agent has chosen and committed to, and thus that is guiding the actions of an agent [17].

Being able to infer intentions from observing opaque agents is crucial in achieving trustworthiness. For instance, in scenarios where agents are deployed in critical decision-making processes, such as autonomous driving, inferring intentions can help anticipate or partial trajectories of actions. For this reason, intentions can be used to build mechanisms for explaining sequences of atomic actions in terms of a common goal.

Furthermore, in cooperative socio-technical systems where humans and AI-based systems coexist, comprehending the intentions behind actions can help the actors align expectations about the behaviour of the other agents and, therefore, collaborate or compete better.

4. Intentions and value alignment

We have mentioned earlier that explanations require a shared code between emitter and receiver, between agent and explainee. This code could be a shared ontology or a shared context. Cultural factors that influence the generation and interpretation processes are usually involved, and such factors can sometimes be related to (human) values, needs, norms, or conventions.

The term value has multiple definitions. In the context of this paper, we use as its definition *the assessment or evaluation of a state of the world in terms of different criteria linked to those values* [18]. It is worth noting that this definition is not exempt from valid criticism [19], but this raises the fact that its definition is still open and requires considering with special care where and *how* exactly are values going to be used. As an example, [20] uses Schwartz's theory of basic values [21] to produce value trees that allow argumentation to decide and motivate which option is preferable from a value-based perspective.

Value alignment is a desirable feature in an AI-based system to achieve trustworthiness, esp. when it is expected that the system behaves in an *ethically-correct* way. When two or more agents align their values, or when a human tries to determine whether an agent is aligned with their values, understanding the intention behind the agent's actions is a fundamental step. After all, empirical studies support the idea that values (such as moral norms) impact the intentions humans commit to [22,23].

To illustrate these concepts, let us return to the autonomous vehicle example. If an AI-based system controlling the vehicle *chooses* an action trajectory that seemingly (through observation) prioritises speed over safety (*e.g.* not slowing down enough in a busy area), understanding this intention is vital because we might be in a case of value misalignment between the designed system and the users of the environment.

A consequence of this line of thought is that traces of actions observed that are interpreted as being the result of an intention can further be re-interpreted, depending on these cultural factors and the values and needs that each explainee holds as their own [24].

5. Can we genuinely infer intentions from traces of opaque agents?

The problem of intent recognition [25], which involves discerning higher-level explanations for an agent's observable actions toward achieving a goal, is still considered an open problem when dealing with opaque agents from mere observations [26].

Some methods proposed in the literature offer solutions to this problem when there is direct access to the agents' plans to be explained [27]. For example, [28] leverages online plan generation within continuous domains, outperforming library-dependent methods but requiring a task-specific planner. Another paradigmatic example is [29], in which an automated pilot models an opponent by tracing their steps within a pre-structured problem-space hierarchy embedded within the SOAR architecture to infer goals through an imitative process. However, these approaches require previous knowledge of some components of these plans, such as tasks or planning domains and are still at a different level of abstraction to intentions.

Another family of methods proposes recognising intentions when the agent's decision process is known and explicit, *i.e* when a partially observable Markov decision process (POMDP) is available. Examples of such methods are found in [30–32].

Some works are already working on preliminary methods to achieve intention recognition from mere observation of agents with a certain degree of success. Case-based approaches [33] can reduce the dependency on task-specific preparations and introduce flexibility in intention recognition. [34] proposes analogical inference and tests it in a controlled *stag-hunt* scenario. [35] explores intention recognition from partial observations in the domain of game-theoretic scenarios.

In summary, as it has yet to be solved, it is uncertain if finding a solution for intention recognition from observing opaque agents is feasible. However, there are some promising works in the literature that, although focused on specific parts of the intention recognition process, can be starting points to research the topic. One possible route could be to combine methods already existing for inferring Markov decision processes (or derived formal isms) from observations [36–38] with methods as mentioned earlier to produce intentions from POMDPs.

6. Conclusions

This paper has proposed making intentions as first-class components in explaining the decision-making process of artificial agents, regardless of whether they are opaque. Also, we have emphasised the critical nature of interpreting such intentions to align agents' actions with human values.

The possibility of being able to infer intentions from mere external observations of agents operating in an environment can provide the basis for developing novel mechanisms for building explanations and helping predict the behaviour of agents. This aspect can be relevant in any socio-technical system where (human or AI) agents have to cooperate and compete.

We have also identified some previous work that can be analysed, combined, and improved to build mechanisms able to infer such intentions. In §5, we argue that this may be a relevant pathway for future research.

Author Contributions: Conceptualization, V.G.-A., L.O.-F., S.A.-N.; formal analysis, V.G.-A., L.O.-F., S.A.-N.; investigation, V.G.-A., L.O.-F., S.A.-N.; supervision, S.A.-N.; writing – original draft, S.A.-N.; writing – review & editing, V.G.-A., L.O.-F., J.V.-S., U.C., S.A.-N.; project administration, J.V.-S., U.C., S.A.-N.; funding acquisition, U.C., S.A.-N. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the AI4CCAM (Grant agreement ID: 101076911) and the HumanE-AI-Net (Grant agreement ID: 952026) European projects.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

References

1. Balke, T.; Gilbert, N. How do agents make decisions? A survey. *Journal of Artificial Societies and Social Simulation* **2014**, *17*, 13.
2. Minsky, M. *Society of mind*; Simon and Schuster, 1988.
3. Malle, B.F. Attribution theories: How people make sense of behavior. *Theories in social psychology* **2011**, *23*, 72–95. Publisher: New York.
4. Rao, A.S.; Georgeff, M.P. Modeling Rational Agents within a BDI-Architecture. Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning; Allen, J.; Fikes, R.; Sandewall, E., Eds. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1991, pp. 473–484.
5. Shvo, M.; Klassen, T.Q.; Sohrabi, S.; McIlraith, S.A. Epistemic plan recognition. Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, 2020, pp. 1251–1259.
6. Broome, J. Are intentions reasons? And how should we cope with incommensurable values. *Practical rationality and preference: Essays for David Gauthier* **2001**, pp. 98–120. Publisher: Cambridge University Press Cambridge.
7. Baker, C.; Saxe, R.; Tenenbaum, J. Bayesian theory of mind: Modeling joint belief-desire attribution. Proceedings of the annual meeting of the cognitive science society, 2011, Vol. 33. Issue: 33.
8. Schuwerk, T.; Kampis, D.; Baillargeon, R.; Biro, S.; Bohn, M.; Byers-Heinlein, K.; Dörrenberg, S.; Fisher, C.; Franchin, L.; Fulcher, T.; others. Action anticipation based on an agent's epistemic state in toddlers and adults. *PsyArXiv* **2021**.
9. Saxe, R.; Baron-Cohen, S. The neuroscience of theory of mind, 2006.
10. Richardson, H.; Saxe, R. Early signatures of and developmental change in brain regions for theory of mind. In *Neural circuit and cognitive development*; Elsevier, 2020; pp. 467–484.
11. Baker, C.L.; Jara-Ettinger, J.; Saxe, R.; Tenenbaum, J.B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* **2017**, *1*, 0064. doi:10.1038/s41562-017-0064.
12. Ho, M.K.; Saxe, R.; Cushman, F. Planning with Theory of Mind. *Trends in Cognitive Sciences* **2022**, *26*, 959–971. doi:10.1016/j.tics.2022.08.003.
13. Rao, A.S.; Georgeff, M.P.; others. BDI agents: from theory to practice. Icmas, 1995, Vol. 95, pp. 312–319.

14. Winikoff, M.; Sidorenko, G. Evaluating a Mechanism for Explaining BDI Agent Behaviour. Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, 2023; AAMAS '23, pp. 2283–2285. event-place: London, United Kingdom.
15. Grice, H.P. Logic and conversation. In *Speech acts*; Brill, 1975; pp. 41–58.
16. Bratman, M. *Intention, Plans, and Practical Reason*; Cambridge, MA: Harvard University Press: Cambridge, 1987.
17. Cohen, P.R.; Levesque, H.J. Intention is choice with commitment. *Artificial intelligence* **1990**, *42*, 213–261. Publisher: Elsevier.
18. Cranefield, S.; Winikoff, M.; Dignum, V.; Dignum, F. No Pizza for You: Value-based Plan Selection in BDI Agents. IJCAI, 2017, pp. 178–184.
19. Van de Poel, I. Embedding values in artificial intelligence (AI) systems. *Minds and Machines* **2020**, *30*, 385–409. Publisher: Springer.
20. van der Weide, T.L.; others. Arguing to motivate decisions. PhD Thesis, Utrecht University, 2011.
21. Schwartz, S.H. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* **2012**, *2*, 11.
22. Godin, G.; Conner, M.; Sheeran, P. Bridging the intention-behaviour gap: The role of moral norm. *British Journal of Social Psychology* **2005**, *44*, 497–512. doi:10.1348/014466604X17452.
23. Young, L.; Saxe, R. When ignorance is no excuse: Different roles for intent across moral domains. *Cognition* **2011**, *120*, 202–214. doi:10.1016/j.cognition.2011.04.005.
24. Yoder, K.J.; Decety, J. The neuroscience of morality and social decision-making. *Psychology, Crime & Law* **2018**, *24*, 279–295. doi:10.1080/1068316X.2017.1414817.
25. Kautz, H.A.; Allen, J.F.; others. Generalized plan recognition. AAAI. Philadelphia, PA, 1986, Vol. 86, p. 5. Issue: 3237.
26. Albrecht, S.V.; Stone, P. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* **2018**, *258*, 66–95. doi:10.1016/j.artint.2018.01.002.
27. Sukthankar, G.; Geib, C.; Bui, H.H.; Pynadath, D.; Goldman, R.P. *Plan, activity, and intent recognition: Theory and practice*; Newnes, 2014.
28. Vered, M.; Kaminka, G.A.; Biham, S. Online goal recognition through mirroring: Humans and agents. The Fourth Annual Conference on Advances in Cognitive Systems, 2016, Vol. 4.
29. Tambe, M.; Rosenbloom, P.S. Event tracking in a dynamic multiagent environment. *Computational Intelligence* **1996**, *12*, 499–522. Publisher: Wiley Online Library.
30. Taha, T.; Miró, J.V.; Dissanayake, G. POMDP-based long-term user intention prediction for wheelchair navigation. 2008 IEEE International Conference on Robotics and Automation. IEEE, 2008, pp. 3920–3925.
31. Bandyopadhyay, T.; Won, K.S.; Frazzoli, E.; Hsu, D.; Lee, W.S.; Rus, D. Intention-aware motion planning. Algorithmic Foundations of Robotics X: Proceedings of the Tenth Workshop on the Algorithmic Foundations of Robotics. Springer, 2013, pp. 475–491.
32. Ramırez, M.; Geffner, H. Goal recognition over POMDPs: Inferring the intention of a POMDP agent. IJCAI. IJCAI/AAAI, 2011, pp. 2009–2014.
33. Kerkez, B.; Cox, M.T. Incremental case-based plan recognition with local predictions. *International Journal on Artificial Intelligence Tools* **2003**, *12*, 413–463. Publisher: World Scientific.
34. Rabkina, I.; Forbus, K.D. Analogical reasoning for intent recognition and action prediction in multi-agent systems. Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems. Cognitive Systems Foundation Cambridge, 2019, pp. 504–517.
35. Markovitch, S.; Reger, R. Learning and exploiting relative weaknesses of opponent agents. *Autonomous Agents and Multi-Agent Systems* **2005**, *10*, 103–130. Publisher: Springer.
36. Hayes, B.; Shah, J.A. Improving Robot Controller Transparency Through Autonomous Policy Explanation. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction; ACM: Vienna Austria, 2017; pp. 303–312. doi:10.1145/2909824.3020233.

37. Liu, T.; McCalmon, J.; Le, T.; Rahman, M.A.; Lee, D.; Alqahtani, S. A novel policy-graph approach with natural language and counterfactual abstractions for explaining reinforcement learning agents. *Autonomous Agents and Multi-Agent Systems* **2023**, *37*, 34. doi:10.1007/s10458-023-09615-8.

38. Domènech i Vila, M.; Gnatyshak, D.; Tormos, A.; Alvarez-Napagao, S. Testing Reinforcement Learning Explainability Methods in a Multi-Agent Cooperative Environment. In *Frontiers in Artificial Intelligence and Applications*; Cortés, A.; Grimaldo, F.; Flaminio, T., Eds.; IOS Press, 2022. doi:10.3233/FAIA220358.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.