When to explain: Identifying explanation triggers in human-agent interaction

Lea Krause and **Piek Vossen**
Computational Lexicology & Terminology Lab (CLTL)
Vrije Universiteit Amsterdam
{l.krause,p.t.j.m.vossen}@vu.nl

Abstract

With more agents deployed than ever, users need to be able to interact and cooperate with them in an effective and comfortable manner. Explanations have been shown to increase the understanding and trust of a user in human-agent interaction. There have been numerous studies investigating this effect, but they rely on the user explicitly requesting an explanation. We propose a first overview of when an explanation should be triggered and show that there are many instances that would be missed if the agent solely relies on direct questions. For this, we differentiate between direct triggers such as commands or questions and introduce indirect triggers like confusion or uncertainty detection.

1 Introduction

The introduction of artificial agents into our daily lives means that an increasing number of lay users interact with them, often even collaborating. This has major societal implications since they are used in domains ranging from healthcare over finance to the military. As a result, special care must be taken to ensure that users understand agents' decisions, can effectively collaborate with them, and even hold them accountable if necessary.

While research emphasising the need for explanations is not new (Buchanan and Shortliffe, 1984), interest has picked up over the past few years (Anjomshoae et al., 2019). Recent advances in artificial intelligence and machine learning have led to a rapid increase in quality of artificial agents. Since most state-of-the-art models are black boxes, it is often not clear to the end-user why the agent made certain decisions. Trust, however, relies on understanding the decision-making of the agent (Lee and Moray, 1992) and trust is a prerequisite for successful collaboration and use. Explanations have been shown to increase the understanding of the agent in human-agent teams (Dzindolet et al., 2003; Wang et al., 2016) and thus increase trust. Within human-human interaction, people resolve conflicts or uncertainties by explaining the reasoning behind their arguments or decisions. Users have a tendency to anthropomorphise agents (Lemaignan et al., 2014) and expect them to behave human-like; thus, they expect them to give explanations for their decisions and actions.

Most work assumes that the user directly asks for an explanation (Sridharan and Meadows, 2019; Ray et al., 2019). We claim that there are many situations where explanations are needed, even if not explicitly requested by the user. In our work, we aim to provide an overview of direct as well as indirect explanation triggers. This overview will be the basis of designing future system experiments and evaluation metrics that target explanations to those needs.

While our primary goal is to investigate this in the context of human-robot interaction, we believe that the impact of these findings is not limited solely to this domain.

2 Related work

In this section, we review recent papers covering explainability, explanations and explanations specifically for human-agent interaction. As our focus lies on human-agent interaction we will mostly refer the reader to survey papers for the parts on explainability and explanations as they give a much more in-depth overview than what would be possible within this space.

2.1 Explainability

Recent years have seen the fundamental expansion of machine learning techniques starting within academia and spreading across industries. While these black-box models bring state-of-the-art results across domains, they are criticised for their

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2020), pages 55–60,
Dublin, Ireland, 18 December 2020.

biases and lack of transparency. The rapid rise of black-box models has resulted in a simultaneous surge of explainability methods. These methods aim to increase the transparency of the models and to make them explainable to humans. Going as far as to include "the right to explanation" in the European Union General Data Protection Regulation (GDPR) (noa, 2016). Adadi and Berrada (2018) have broken the need for explainable artificial intelligence down into four reasons: explain to justify, explain to control, explain to improve and explain to discover. The last two especially show that explainability does not need to slow a model down, but can instead further its development and share new discoveries it has made.

Although there has been a large number of publications in explainable artificial intelligence in recent years, no common taxonomy or agreed meaning has emerged. Two recent in depth proposals were done by Lipton (2016) and Sokol and Flach (2020). The latter propose a fact sheet detailing five dimensions to guide the development of future explainability approaches: 1. functional requirements, 2. operational requirements, 3. usability criteria 4. security, privacy and any vulnerabilities, 5. validation. Their approach is one of the few taking results from other disciplines, such as sociology and psychology, into account, which have been studying explainability and explanations much longer than artificial intelligence.

This lack of consideration of input from other disciplines is the topic of a thorough critique of the current state of explainable artificial intelligence by Mittelstadt et al. (2019). They examine the discrepancy between what designers and end-users want from explanations and come to the conclusion that explanations as they currently exist in artificial intelligence fail in providing adequate explanations to those affected by the results of the machine learning algorithms. Their recommendations to resolve this discrepancy are based on Miller (2019) whose findings we will discuss in the next paragraph.

2.2 Explanations

Explanations differ from general explainability in that they focus only on explaining a single prediction instance of a model or in our case, agent. The most extensive review of explanations within A.I. in recent years has been done by Miller (2019). He reviews existing research on explanations from social sciences, philosophy, psychology and cognitive science, and connects it to the current discourse in explainable artificial intelligence. His main conclusion is that explanations need to be contextualised instead of just stating a causal relation. He breaks this down into four findings:

1. An explanation should be contrastive, they are an answer to the question *Why did **A** happen instead of **B**?*

2. The selection of an explanation is biased; selected causes are chosen to fit the explanation

3. A probability alone does not make an explanation.

4. An explanation is part of a social interaction, related to the mental states of the participants of the conversation.

2.3 Human-agent interaction

Explanations for human-agent interaction often form a challenging task. They have to be generated in different circumstances with somewhat unpredictable input (unpredictable humans) and most people the agent will interact with are not experts, therefore the explanations have to be understandable for a lay-person.

Anjomshoae et al. (2019) have conducted a large-scale literature review on current literature (after 2008) on explainable agents and robots. Similarly to the field of explainability in general, they have found a rapid increase in works published since 2016. The similarities continue, as only 37% of the papers made any reference to the theoretical background of explanations. The main direction found to be relevant for future work is the communication of the explanations.

Rosenfeld and Richardson (2019) propose a taxonomy for explainability in human-agent systems in which they cover the questions of: *Why* is there a need for explainability?, *Who* is the target audience?, *What* kind of explanation should be generated? *When* should the explanation be presented to the human? and lastly *How* can the explanations be evaluated?

Another overview from a different angle was done by Sridharan and Meadows (2019). While they as well give a framework for explanations in human-robot collaboration, their main contribution is their investigation of combining knowledge representation, reasoning, and learning to generate interactive explanations.

Several other studies have investigated the effectiveness of explanations for task-oriented human-agent teams and reported an increased success rate and self-reported trust in the agent (Ray et al., 2019; Chakraborti et al., 2019; Gong and Zhang, 2018; Wang et al., 2016)

Recently, post-hoc explanations have been accused of fairwashing (Aïvodji et al., 2019) and Rudin (2019) specifically called for researchers to focus on completely interpretable models if it is a high stakes decision. Agents can be deployed in many circumstances, also high stake ones. We agree that only post-hoc explanations of blackbox models are not enough under these circumstances, but we believe that explanations are nevertheless important in the case of human-agent interaction as they fulfil a communicative function as well as an informative one.

3 Triggers

All the work on explanations for human-robot agents mentioned before makes the assumption that the user is explicitly going to ask for an explanation and to the best of our knowledge, the question when during communication an explanation is actually needed remains unanswered. Rosenfeld and Richardson (2019) pose the question in their overview paper on explainability in human-agent systems, but only consider the task of the agent-system, not the communicative aspect or any flexible trigger detection that takes the users current state into account. We argue that it is vital to fill this gap in order to make use of the full potential of explanations for human-agent interaction, as there are many situations in which an explanation is needed, even if not explicitly requested by the user. We therefore provide a first overview of possible direct and indirect triggers.

When users interact with explainable agents, the agent constantly has to evaluate whether it has to inform the user about its decisions. To do this efficiently it needs clear triggers when to explain.

3.1 Direct triggers

The most obvious triggers of explanations are explicitly expressed **commands or questions**. According to Miller (2019), an explanation is inherently an answer to a why-question. There are different underlying causes for such an explicit question. As described earlier, *trust* plays a significant role in human-agent interaction. One of the principal reasons for the user to demand an explanation is thus when they mistrust the decision of the agent and need clarification. Secondly, the user could be uncertain whether they have understood the agent correctly and seek an explanation to resolve this *uncertainty*. One step further is the occurrence of a *knowledge gap*. Here, the user might be completely unfamiliar with the topic of a decision. This case is also a critical one, as the user otherwise could not judge whether the decision is correct. Consequently, the reliability of the explanation is likewise of utmost importance. Lastly, it could simply be *curiosity*, due to interacting with a new agent or a topic that sparked the interest of the user. These examples also show that the agent should tailor its explanations to the underlying trigger.

Direct triggers	**Indirect triggers**
Command / Question	Confusion detection
Urgency	Agent uncertainty
	Conflicting mental states
	Conflict of interest
	Lack of trust

Table 1: Overview of direct and indirect triggers of explanations

We argue that additionally there are cases where an inherent **urgency** to inform is inherent to the topic without the occurrence of a question or uncertainty. This case is particularly relevant in the context of agents. The agent might observe something the human has overlooked. A situation like this would rely heavily on the agent's reasoning capabilities. It needs to analyse the situational urgency, the potential impact, and react instantaneously. Common use cases for this are agents deployed in elderly homes.

3.2 Indirect triggers

The often multi-modal nature of agents gives the opportunity to detect the need for an explanation not solely by relying on explicit commands. Indirect triggers largely depend on signal interpretation and belief detection. An example from educational computing is confusion detection (Arguel et al., 2017; Bosch et al., 2014). Detecting confusion is an especially fitting case for explainable agents, as they can also use it to detect whether an explanation was successful. Firstly, the agent can use **visual cues**, here we draw from Arguel et al. (2017) findings for detecting confusion in digital learning

environments. This can be *eye-tracking*, where the user's gaze is captured. Direction and duration of the user's eye movement can indicate their focus of attention as well as their emotional status. Eye-tracking is not deemed suitable for online learners, as to not add extra equipment. Agents, however, are often equipped with high-resolution cameras and object recognition, making them suitable for this type of detection. Another visual cue are *facial expressions*. Facial expressions have long been used in affect detection. Lowering the eyebrows paired with tightening the eyelids are indicators of confusion (D'Mello et al., 2009). *Body posture and movement* are further indicators. These can include shoulder position, hand placement and movements like head-scratching.

A second possible modality are **audio cues**. In *prosody*, rising intonation can indicate uncertainty and is more often paired with a wrong answer to a question than falling intonation (Brennan and Williams, 1995). *Speech disfluency*, like filler words such as "huh", "uh" or "um", occur more often if the speaker is uncertain or is presented with a choice. There is even a hierarchy as "um" marks a a greater uncertainty than "uh" (Brennan and Williams, 1995; Corley and Stewart, 2008).

An important step towards transparency is, that if the robot detects an **uncertainty**, be it from an unclear signal or the occurrence of multiple equally likely solutions, it gives an explanation why the decision might not be trustworthy.

The following triggers prompt explanations used for **reconciliation** between the agent and the user.

A more abstract trigger for an explanation can be found within **theory of mind** (Shvo et al., 2020; Miller, 2019). If the agent detects conflicting beliefs or mental states between itself and the user, it can take the user's beliefs into account and try to resolve them. These conflicting beliefs can be many fold, for one it can be distinguished between a misunderstanding and a misconception (McRoy and Hirst, 1995). A *misunderstanding* occurs when one side does not succeed in conveying the beliefs that they wanted to convey to their conversational partner. This is for example the case if a student misunderstands the question on an exam (Olde Bekkink et al., 2016). *Misconceptions* on the other hand are related to factual states of the world. Here the user could have an incorrect belief about what is a case in the world or what could be a case in the world (Webber and Mays, 1983), for instance believing that Northern Ireland is part of Great Britain.

The last potential trigger is a **conflict of interest**. In this case, there is a full understanding between the agent and the human, but a disagreement about the planning or the method to reach the goal. The agent needs to explain itself to either reach an agreement or for the user to be able to make an informed choice to disregard the agent's suggestion.

While we have described the triggers as separate entities, users will benefit most, if all of the signals are processed simultaneously by the agent.

4 Conclusion

We have shown the need for detecting triggers of explanations and given a first classification of possible internal and external triggers. Next steps will be to implement this classification system into an agent. Further work will then correlate the triggers to specific types of explanations and their generation.

5 Acknowledgments

This research was funded by the Vrije Universiteit Amsterdam, the Netherlands Organisation for Scientific Research via the Spinoza grant awarded to Piek Vossen and the Hybrid Intelligence Centre via the Zwaartekracht grant from the Dutch Ministry of Education, Culture and Science.

References

2016. General Data Protection Regulation (GDPR) – Official Legal Text.

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160. Conference Name: IEEE Access.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Amaël Arguel, Lori Lockyer, Ottmar V. Lipp, Jason M. Lodge, and Gregor Kennedy. 2017. Inside Out: Detecting Learners' Confusion to Improve Interactive Digital Learning Environments. *Journal of Educational Computing Research*, 55(4):526–551. Publisher: SAGE Publications Inc.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp.

2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36 th International Conference on Machine Learning,*, page 10, Long Beach, California.

Nigel Bosch, Yuxuan Chen, and Sidney D'Mello. 2014. It's Written on Your Face: Detecting Affective States from Facial Expressions while Learning Computer Programming. In *Intelligent Tutoring Systems*, pages 39–44, Cham. Springer International Publishing.

Susan E Brennan and Maurice Williams. 1995. The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34(3):383–398. Publisher: Elsevier.

Bruce G Buchanan and Edward H Shortliffe. 1984. Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley series in artificial intelligence). Publisher: Addison-Wesley Longman Publishing Co., Inc.

Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan Explanations as Model Reconciliation – An Empirical Study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 258–266. ISSN: 2167-2148.

Martin Corley and Oliver W. Stewart. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Sidney D'Mello, Scotty Craig, and Arthur Graesser. 2009. Multi-method assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, 4(3-4):165–187.

Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718.

Ze Gong and Yu Zhang. 2018. Behavior Explanation as Intention Signaling in Human-Robot Teaming. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1005–1011. ISSN: 1944-9437.

John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270. Publisher: Taylor & Francis.

Séverin Lemaignan, Julia Fink, and Pierre Dillenbourg. 2014. The Dynamics of Anthropomorphism in Robotics. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 226–227. ISSN: 2167-2121.

Zachary C. Lipton. 2016. The Mythos of Model Interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning*, New York, NY, USA.

Susan W. McRoy and Graeme Hirst. 1995. The Repair of Speech Act Misunderstandings by Abductive Inference. *Computational Linguistics*, 21(4):435–478.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38. Publisher: Elsevier.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 279–288. ArXiv: 1811.01439.

Marleen Olde Bekkink, A. R. T. Rogier Donders, Jan G. Kooloos, Rob M. W. de Waal, and Dirk J. Ruiter. 2016. Uncovering students' misconceptions by assessment of their written questions. *BMC Medical Education*, 16(1):221.

Arijit Ray, Yi Yao, Rakesh Kumar, Ajay Divakaran, and Giedrius Burachas. 2019. Can you explain that? Lucid explanations help human-AI collaborative image retrieval. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 153–161. Issue: 1.

Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. Publisher: Nature Publishing Group.

Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith. 2020. Towards the Role of Theory of Mind in Explanation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 75–93, Cham. Springer International Publishing.

Kacper Sokol and Peter Flach. 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67. ArXiv: 1912.05100.

Mohan Sridharan and Ben Meadows. 2019. Towards a Theory of Explanations for Human–Robot Collaboration. *KI - Künstliche Intelligenz*, 33(4):331–342.

Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–116. ISSN: 2167-2148.

Bonnie Lynn Webber and Eric Mays. 1983. Varieties of user misconceptions: detection and correction. In *Proceedings of the Eighth international joint conference on Artificial intelligence - Volume 2*, IJCAI'83, pages 650–652, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.