Towards the Role of Theory of Mind in Explanation (Extended Abstract)*

Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith

Department of Computer Science, University of Toronto, Toronto, Canada
Vector Institute, Toronto, Canada
{maayanshvo,toryn,sheila}@cs.toronto.edu

Abstract. Theory of Mind is commonly defined as the ability to attribute mental states (e.g., beliefs, goals) to oneself, and to others. A large body of previous work—from the social sciences to artificial intelligence—has observed that Theory of Mind capabilities are central to providing an explanation to another agent or when explaining that agent's behaviour. In this paper, we build and expand upon previous work by providing an account of explanation in terms of the beliefs of agents and the mechanism by which agents revise their beliefs given possible explanations. We further identify a set of desiderata for explanations that utilize Theory of Mind. These desiderata inform our belief-based account of explanation.

1 Introduction

Following Premack and Woodruff [22], an agent exercises *Theory of Mind* if it imputes mental states to itself and others. Here we explore the role of Theory of Mind in explanation. Consider the following narrative by way of illustration.

> *Mary, Bob and Tom are housemates sharing a house. While Tom was away on a business trip, Mary and Bob noticed a hole in the roof of their house and called a handyman to fix it. Before the handyman could come, however, it rained during the night and the floor got wet. Bob, who sleeps in a windowless room, did not notice the rain. Tom, who just got back from his trip that day, noticed the rain but did not know about the hole in the roof. Mary saw Tom return to the house at night and so knew that Tom knew that it had rained. In the morning, when trying to explain the wet floor to Bob, Mary tells him that it had rained during the night and when explaining to Tom she tells him that she and Bob had discovered a hole in the roof (adding that the handyman will arrive the next day).*

Clearly, Mary tailored her explanations to each of her housemates, believing the information she was providing to them was sufficient to explain the wet floor in their respective mental states. Her ability to do this stems from her

* The full paper [25] was presented at EXTRAAMAS 2020 and can be found here: https://link.springer.com/chapter/10.1007/978-3-030-51924-7_5

Theory of Mind - her ability to attribute mental states (e.g., beliefs) to herself and to others. In humans, the use of Theory of Mind in explanation has been demonstrated empirically by Slugoski et al. [26] via a set of experiments where human participants gave different explanations to different explainees (i.e., the recipient of an explanation), based on the beliefs of the explainers about the beliefs of the explainees[1]. Of course Mary's explanations are only as good as her ability to model the mental states of her housemates and how they will alter their mental states in light of her explanation. Mary's beliefs about Bob and Tom's beliefs, or her belief about how each of them revises their beliefs, may well be wrong, in which case her explanations to them may fail to explain why the floor is wet.

Explanation has been studied in a diversity of disciplines. Miller [17] provides an extensive survey of explanation in artificial intelligence that includes a selection of historical works in philosophy (e.g., Hempel and Oppenheim [13]; Peirce [19]; Harman [11]), arguing for the important role of philosophy and the social sciences in future work on explanation. Within AI, early work on explanation included a variety of logic-based and probabilistic approaches to abductive inference or so-called *inference to the best explanation* including the early works of Pople [21], Charniak and McDermott [6], Poole [20], and Levesque [15]. In the mid 1980s, explanation was popularized in the context of expert systems where explanations were often generated by backward chaining over a set of symbolic inference steps (e.g., [12, 24]). Following that time, explanation was a common element in a diversity of applications of symbolic AI reasoning (e.g., [16, 1, 27]). The recent resurgence of interest in explanation is largely in the guise of so-called *Explainable AI* (XAI), which is motivated by the need to provide human-interpretable explanations for decision making in black-box classification and decision-making systems based on machine and deep learning (e.g., Samek et al. [23]; Gunning et al. [8]).

Numerous researchers have acknowledged the importance of Theory of Mind in explanation. In the 80s and 90s, formal accounts of explanation such as those proposed by Gärdenfors [7] and Chajewska and Halpern [4] observed that an explanation for one agent may not serve as an explanation for another, and the explainer must therefore tailor an explanation to an explainee given the latter's beliefs. Within the space of user modelling and dialogue, and also set in the 80s and 90s, Weiner's [29] BLAH system and Cawsey's [3] EDGE system both tailor explanations to the presumed user model. More recently, researchers have leveraged belief-desire-intention (BDI) architectures as a natural framework for explanations reflecting Theory of Mind. Such software architectures can enable an explainer to explicitly represent its own beliefs, desires, and intentions, as well as those of an explainee, and to relate explanations to its own beliefs and goals or those of the explainee (e.g., Harbers et al. [10]; Kaptein et al. [14]). Most recently, Westberg et al. [30] has posited that incorporating various points of view on Theory of Mind from the cognitive sciences will facilitate the creation of

[1] We henceforth use *explainer* and *explainee* in reference to the provider and recipient of the explanation, and *explanandum* in reference to the thing to be explained.

agents better suited to communicate and explain themselves to the humans with whom they are interacting. Additionally, Miller [17] has surveyed this body of work and has also emphasized the importance of the explainer's ability to tailor an explanation to the explainee, using its understanding of the latter's mind. Finally, within the subfield of XAI known as XAI Planning (XAIP) Chakraborti et al. [5] have implemented XAIP in human-agent teaming settings, such as search & rescue, where a robot equipped with Theory of Mind capabilities could explain its actions to its human teammate by taking into account the latter's mental state.

2 Synopsis

The use of Theory of Mind in explanation holds the promise of producing high-quality explanations that are tailored to the beliefs of the explainee, in the context of the beliefs (and ignorance) of the explainer. In this paper, we identify a set of desiderata for explanation that utilizes Theory of Mind. Our work was strongly influenced by the recognition that explainers and explainees can take on many different forms—human or machine—and that their beliefs may be internally represented, inspected, and revised in diverse ways. For example, the agent's beliefs may be stored in a human brain or in, for instance, the weights of a neural network or formulae in a knowledge base and, the extent of those beliefs may be limited by the reasoning capabilities of the agent. Moreover, we posit that an account of explanation must model the possibly false or simply incomplete beliefs of explainers and explainees, and allow an explainer to reason about the explainee's beliefs when providing the latter with an explanation. This is crucial since, due to their possibly differing beliefs, an explanation for the explainer may not be an explanation for the explainee.

The main contribution of our work is a belief-based account of explanation that satisfies all of the aforementioned desiderata by employing a number of crucial building blocks. Namely, in order to capture the diversity of human and machine explainers and explainees, our account of explanation employs epistemic states to capture the mental states of *both* the explainer and explainee, and incorporates a belief revision operator to assimilate explanations into the explainee's epistemic state. Our account finds its origins in works that attributed agents with mental states in the form of epistemic states (with seminal work by Gärdenfors [7] and later notable work by Levesque [15]; Boutilier and Becher [2]; Chajewska and Halpern [4]; and Halpern and Pearl [9]).

Further, we discuss the criteria by which the quality of an explanation can be evaluated. Moreover, we formalize and discuss the notion of a discrepancy as a property that allows the explainer to anticipate and provide explanations without prompting. We also discuss properties relating to the adequacy of the explainer's beliefs, exploring when the explainer's beliefs about the explainee's beliefs and reasoning capabilities are accurate 'enough' for the explainer to generate 'good' explanations wrt the explainee.

This paper provides a general characterization of explanation without focusing on its computational realization. This is done by design to allow for a diversity of explanation scenarios and agent types, including human, black-box decision maker, or knowledge-based system. Nevertheless in the simplest case if the beliefs of the explainer are represented as formulae (logical or probabilistic) then, as observed by Levesque [15] and Boutilier and Becher [2], our notion of explanation may be realized via an augmentation of existing abductive reasoning systems such as Theorist (Poole [20]), for example.

In our work, we mostly relate our Theory of Mind characterization of explanation in the context of English-like statements (e.g., Mary *telling* Bob that it had rained last night). However, if we turn to the broad endeavour of XAI that helped motivate our account, we note that an explanation can take on many different forms other than human-interpretable language (e.g., a set of weights in a neural network, select pixels, a gesture, a heightening of intensity in a region of an image). At its core, an explanation is something that is conveyed by the explainer to the explainee (e.g., by telling, demonstrating, visualizing, etc.) in order to justify the latter's belief in some explanandum. For example, by constructing a heat-map from a medical image, an otherwise black-box decision-making algorithm can highlight for the explainee the pixels that have most strongly supported its classification decision [18], thereby allowing the explainee to assimiliate this explanation into their beliefs and better interpret the system's decision. As we argue in our work, the decision-making system, acting as an explainer, should possess the ability to take the epistemic state of the explainee into account, tease apart the salient features required for the explainee to justify its belief in the explanandum, and present those to the explainee as an explanation. Some of these insights pertaining to explanations for black-box solvers are similarly echoed by Sreedharan et al. in the context of their model reconciliation paradigm [28, Section 2]. Our general account is intended to provide building blocks towards this broader XAI objective.

There are several take-aways from this paper that are worth highlighting. Explanations need not be consistent with an agent's beliefs. As such, contrary to most logical treatments of explanation, characterizations of explanation should involve a belief revision component, and not just the expansion of existing beliefs to include an explanation. Further, by providing a belief-based account of explanation that characterizes mental states in terms of epistemic states, and by allowing for epistemic states and revision operators to be realized in a diversity of forms from standard logical accounts, to computer programs, neural networks or human brains, we can capture the mental states of a myriad of different types of agents. Finally, by characterizing explanations in terms of the explainer's beliefs about the explainee's beliefs and revision operator, we can capture the role of Theory of Mind in explanation for a myriad of different types of agents.

Bibliography

[1] Borgida, A., Calvanese, D., Rodriguez-Muro, M.: Explanation in DL-Lite. In: Proceedings of the 21st International Workshop on Description Logics (DL2008). CEUR Workshop Proceedings, vol. 353 (2008)

[2] Boutilier, C., Becher, V.: Abduction as Belief Revision. AIJ 77(1), 43–94 (1995)

[3] Cawsey, A.: Generating interactive explanations. In: AAAI. pp. 86–91 (1991)

[4] Chajewska, U., Halpern, J.Y.: Defining explanation in probabilistic systems. arXiv preprint arXiv:1302.1526 (2013)

[5] Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In: IJCAI. pp. 156–163 (2017)

[6] Charniak, E., McDermott, D.: Introduction to Artificial Intelligence. Addison Wesley (1985)

[7] Gärdenfors, P.: Knowledge in flux: Modeling the dynamics of epistemic states. The MIT press (1988)

[8] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.: XAI - Explainable Artificial Intelligence. Sci. Robotics 4(37) (2019)

[9] Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part ii: Explanations. The British journal for the philosophy of science 56(4), 889–911 (2005)

[10] Harbers, M., Van den Bosch, K., Meyer, J.J.: Modeling agents with a theory of mind: Theory–theory versus simulation theory. Web Intelligence and Agent Systems: An International Journal 10(3), 331–343 (2012)

[11] Harman, G.H.: The inference to the best explanation. The philosophical review 74(1), 88–95 (1965)

[12] Hayes-Roth, F., Waterman, D.A., Lenat, D.B. (eds.): Building expert systems. Teknowledge Series in Knowledge Engineering, Addison-Wesley (1983)

[13] Hempel, C.G., Oppenheim, P.: Studies in the logic of explanation. Philosophy of science 15(2), 135–175 (1948)

[14] Kaptein, F., Broekens, J., Hindriks, K., Neerincx, M.: Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 676–682. IEEE (2017)

[15] Levesque, H.J.: A Knowledge-Level Account of Abduction. In: IJCAI. pp. 1061–1067 (1989)

[16] McGuinness, D.L., da Silva, P.P.: Explaining answers from the semantic web: the inference web approach. Journal of Web Semantics 1(4), 397–413 (2004)

[17] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. AIJ 267, 1–38 (2019)

[18] Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73, 1–15 (2018)

[19] Peirce, C.: Deduction, induction and hypothesis. Popular Science Monthly 13 (1878)

[20] Poole, D.: Explanation and prediction: an architecture for default and abductive reasoning. Computational Intelligence 5(2), 97–110 (1989)

[21] Pople, H.E.: On the mechanization of abductive logic. In: IJCAI. pp. 147–152 (1973)

[22] Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and brain sciences 1(4), 515–526 (1978)

[23] Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)

[24] Shortliffe, E.H., Buchanan, B.G.: Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley Publishing Company (1985)

[25] Shvo, M., Klassen, T.Q., McIlraith, S.A.: Towards the role of theory of mind in explanation. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems - Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9-13, 2020, Revised Selected Papers. Lecture Notes in Computer Science, vol. 12175, pp. 75–93. Springer (2020)

[26] Slugoski, B.R., Lalljee, M., Lamb, R., Ginsburg, G.P.: Attribution in conversational context: Effect of mutual knowledge on explanation-giving. European Journal of Social Psychology 23(3), 219–238 (1993)

[27] Sohrabi, S., Baier, J.A., McIlraith, S.A.: Preferred explanations: Theory and generation via planning. In: AAAI. pp. 261–267 (2011)

[28] Sreedharan, S., Hernandez, A.O., Mishra, A.P., Kambhampati, S.: Model-free model reconciliation. In: IJCAI. pp. 587–594 (2019)

[29] Weiner, J.: Blah, a system which explains its reasoning. AIJ 15(1-2), 19–48 (1980)

[30] Westberg, M., Zelvelder, A., Najjar, A.: A historical perspective on cognitive science and its influence on XAI research. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 205–219. Springer (2019)