Towards the Role of Theory of Mind in Explanation

Maayan Shvo[1,2](✉), Toryn Q. Klassen[1,2], and Sheila A. McIlraith[1,2]

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

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 Woodru [38], 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 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

D. Calvaresi et al. (Eds.): EXTRAAMAS 2020, LNAI 12175, pp. 75–93, 2020.
https://doi.org/10.1007/978-3-030-51924-7_5

demonstrated empirically by Slugoski et al. [44] 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 [30] provides an extensive survey of explanation in artificial intelligence that includes a selection of historical works in philosophy (e.g., Hempel and Oppenheim [21]; Peirce [34]; Harman [19]), 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 [37], Charniak and McDermott [10], Poole [35], and Levesque [26]. 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., [20,43]). Following that time, explanation was a common element in a diversity of applications of symbolic AI reasoning (e.g., [3,28,45]). 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. [41]; Gunning et al. [16]).

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 [12] and Chajewska and Halpern [7] 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 [49] BLAH system and Cawsey's [6] 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. [18]; Kaptein et al. [24]). Most recently, Westberg et al. [50] has posited that incorporating various points of view on Theory of Mind from the cognitive sciences will facilitate the creation of agents better suited to communicate and explain themselves to the humans with whom they are interacting. Additionally, Miller [30] has surveyed this body 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.

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. [8] 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.

In this paper we build on the shoulders of previous scholarly work to explore the role of Theory of Mind in explanation with a view to addressing the diverse needs of explanation in AI, and XAI in particular. To this end, in Sect. 2 we identify a set of desiderata for explanations that utilize Theory of Mind. These desiderata inform a set of design choices for a belief-based account of explanation which we present in Sect. 3. Of course not all explanations are created equal, and in Sect. 4 we discuss the criteria by which the quality of an explanation can be evaluated. In Sect. 5 we demonstrate how, in the absence of an explicit prompt to be explained, our account allows the explainer to simulate the explainee's mental state and identify discrepancies that warrant explanation. Explanations are limited by the coverage and accuracy of the explainer's beliefs as well as its reasoning capacity. In Sect. 6, we show how our account allows for the modelling of the ignorance and misconceptions of an explainer pertaining to the mental state of an explainee and how these may affect the quality of explanation. We conclude with a discussion of related work and possible computational realizations of our general account.

2 Desiderata for Theory of Mind in Explanation

We begin our investigation by reflecting on the key components that support an agent in imputing mental states to itself and others, reasoning about how the provision of new information is assimilated into an agent's existing set of beliefs, and the circumstances underwhich such information constitutes an explanation for the explainee. To this end, we identify a set of desiderata that inform our account of explanation in the sections to follow.

multi-agent: the account must be conceived in a multi-agent setting to support representation of the beliefs of one or more explainer and explainee.

agent-type agnostic: the account must support a myriad of different agent types whose beliefs may be internally represented, inspectable, and revisable in diverse ways. For example, the agent's beliefs may be stored in a human brain or in, for instance, the parameters of a neural network or formulae in a knowledge base.

belief based: the account must model the possibly false or simply incomplete beliefs of explainers and explainees.

reason about the beliefs of others: the account must allow an explainer to reason about the explainee's beliefs when providing the latter with an explanation since, due to their possibly differing beliefs, an explanation for the explainer may not be an explanation for the explainee.

support belief revision: the account must enable the explainer to consider how an explanation is assimilated by the explainee, and in particular how the latter revises their beliefs given potential explanations which may be inconsistent with their current beliefs.

explanations can refer to beliefs: the account must allow for explanations that themselves refer to beliefs. To illustrate why this is useful, consider that the explainer might explain their having not told the explainee the location of a party by saying that the explainer believed that the explainee knew the location.

While previous work has addressed some of these desiderata, in this paper we propose a belief-based account of explanation in terms of epistemic states of agents that satisfies all of the aforementioned desiderata by employing a number of crucial building blocks relating to these desiderata.

3 A Belief-Based Account of Explanation

We appeal to logics of belief to provide a belief-based account of explanation in the context of Theory of Mind.

Many logical accounts of explanation assume the existence of a knowledge base—a logical axiomatization of the domain in terms of a set of formulae (e.g., [5]). With such a knowledge base in hand, a popular logic-based characterization of explanation is in terms of abduction as follows.

Definition 1 (Abductive Explanation (after [36])). *Given a logical theory, T, and an explanandum O, E explains O given a theory T if $T \cup E \models O$ and $T \cup E$ is consistent.*

Here we make no such commitment to the representation of beliefs in terms of a set of logical formulae. Rather, in order to capture the diversity of human and machine explainers and explainees, our account finds its origins in works that attributed agents with mental states in the form of epistemic states (with seminal work by Gärdenfors [12] and later notable work by Levesque [26]; Boutilier and Becher [4]; Chajewska and Halpern [7]; and Halpern and Pearl [17]).

3.1 Mental States as Epistemic States

We employ the notion of an epistemic state, e, or in the case of multiple agents, a collection of epistemic states, $\vec{e}$, to capture the beliefs of agents. These are used to provide the semantics for the language below.

We will suppose that we have a finite set of agents, $A = \{1, 2, \ldots, n\}$, and a set of propositional symbols P. We define a language

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid B_i\varphi \mid [\varphi]_i\varphi \tag{1}$$

where $p \in P$ and $i \in A$. We introduce $\perp$ as an abbreviation for $(p \wedge \neg p)$ for an arbitrary $p \in P$.

The intended meaning of $B_i\varphi$ is that agent i believes φ, and the intended meaning of $[\alpha]_i\varphi$ is that after agent i revises their beliefs by α, φ is true.

We assume that our epistemic states are such that we can say that a formula φ is true at e when φ is believed. To be clear, although we use formulas to describe what is believed, an epistemic state is not in general *defined* as a set of formulas, nor required to be represented internally as one. For a conventional example, e might be a set of possible worlds with accessibility relations and so on. However, we also allow for epistemic states to take very different forms. For example, one might want to model limited reasoning capabilities in some manner to avoid the so-called problem of logical omniscience [48], in which agents unrealistically believe all the deductive consequences of their beliefs. We might also wish for our epistemic states to be realized in terms of a computer program, such as a neural network, or via a human brain.

Furthermore, we assume we have a *revision operator* $*$ so that $e * \alpha$ is another epistemic state, the result of revising by α. We will use * in defining the semantics for the $[\alpha]_i$ operator. Much as we have not committed to a particular structure for epistemic states, we will not commit to a particular revision operator. A large body of work has studied belief change in agents where belief revision typically concerns belief change in a static environment, possibly in the presence of incorrect and partial beliefs. Amongst the most popular guidelines for belief revision are the AGM postulates [1], and the DP postulates [11] (for iterated revision). We will not require that our $*$ satisfies these properties except where noted. Similarly to the situation with our epistemic states, we might want our revision operator to be realized in terms of a computer program or human reasoning.

While epistemic states assign a truth value to any formula in our language – the language given by the grammar in (1) – that value indicates whether the formula is believed by the agent in question, not whether it's actually true. From an objective point of view, the formulas whose truth values we can determine are from the subset of the language consisting of formulas which are concerned only with beliefs. We define this subset of formulas below:

Definition 2 (Agent Formula). *An agent formula is one in which no atomic symbol appears outside the scope of a belief operator, i.e., a formula ϕ of the form*

$$\phi ::= B_i\varphi \mid \neg\phi \mid (\phi \wedge \phi) \mid [\varphi]_i\phi \tag{2}$$

where φ is any (possibly non-agent) formula.

We assign truth values to agent formulas with a collection of epistemic states $\vec{e} = e_1, \ldots, e_n$ (corresponding to the different agents) according to the satisfaction relation $\models$ below.

- $\vec{e} \models B_i\varphi$ iff φ is true at e_i
- $\vec{e} \models \neg\phi$ iff $\vec{e} \not\models \phi$
- $\vec{e} \models (\phi \wedge \psi)$ iff $\vec{e} \models \phi$ and $\vec{e} \models \psi$
- $\vec{e} \models [\alpha]_i\phi$ iff $\langle e_1, \ldots, (e_i * \alpha), \ldots, e_n\rangle \models \phi$

Note that the semantics of the $[\alpha]_i$ operator is defined using the revision operator.

Give this abstract framework for talking about beliefs, we can define explanations. The lack of commitment to the form of the epistemic state and revision operator is important because it affords us the ability to model a diversity of agents. In so doing, for the definitions of explanation that follow, the explainer will have beliefs about the other agents' beliefs and about their revision operators, and the effectiveness of the explainer's explanations for any particular agent will rely on the fidelity of those beliefs.

3.2 Characterizing Explanations

Definition 3 (Explanation). *Given epistemic states $\vec{e}$, we say that α explains β for agent i if $\vec{e} \models [\alpha]_i(B_i\beta \wedge \neg B_i\bot)$.*

Notation: For notational convenience, we define $\mathit{Expl}(i, \alpha, \beta)$ as an abbreviation for $[\alpha]_i(B_i\beta \wedge \neg B_i\bot)$.

That is, α explains β if revising by α makes agent i believe β while still having consistent beliefs.[2] Note that (with respect to revising by non-modal formulas) if revision of agent i's epistemic state satisfies the AGM postulates, then the result of revision will be inconsistent only if either the agent initially had inconsistent beliefs, or if α itself is inconsistent.

Intuitively, our definition of explanation allows for more explanations than the traditional account in Definition 1. For one thing, we allow explanations to refer to modal operators. Even without that, though, an important difference is that our definition is in terms of belief revision and so allows for an explanation that isn't consistent with the agent's initial beliefs. Our account builds upon prior accounts of explanation defined relative to belief revision such as Boutilier and Becher [4] and Nepomuceno-Fernández et al. [32].

To make the comparison more explicit, consider defining an epistemic state e_i as a propositional theory T, as in the following theorem.

Theorem 1. *Suppose that e_i is defined as being a propositional theory T, and that the formulas e_i makes true are defined to be the logical consequences of T (note that these are restricted to the non-modal subset of our language). Suppose furthermore that the revision operator $*$ on e_i satisfies the AGM postulates (w.r.t. non-modal formulas). Then for non-modal formulas α and β, $\vec{e} \models \mathit{Expl}(i, \alpha, \beta)$ if $T \cup \{\alpha\}$ is consistent and $T \cup \{\alpha\} \models \beta$.*

Proof. Because $T \cup \{\alpha\}$ is consistent, by the AGM "vacuity" postulate, $T * \alpha$ is equal to the expansion of T by α, that is, the closure of $T \cup \{\alpha\}$. Therefore, $T * \alpha \models \beta$.

[2] If agent i is not logically omniscient, requiring i to not believe $\bot$ may not prevent i's beliefs from being inconsistent in some subtler way. For example, i might both believe p and believe $\neg p$, even though it does not believe $(p \wedge \neg p)$.

However, we may also get further explanations. In the circumstances described by Theorem 1, if $T \cup \{\beta\}$ is inconsistent, then Definition 1 would say there are no explanations of β given the theory T, while there may be formulas that agent with epistemic state T can revise by that would make them believe β.

It is also possible to talk in the language about agents' beliefs about $\mathit{Expl}(i, \alpha, \beta)$, i.e. about whether α explains β for agent i.

Definition 4 (Subjective Explanation). *Given epistemic states $\vec{e}$, we say that α explains β for agent j from agent i's perspective, if $\vec{e} \models B_i\mathit{Expl}(j, \alpha, \beta)$.*

Example 1. We formalize our example from Sect. 1. We assume that Mary, Bob and Tom all believe (and believe that the other agents believe) $rain \wedge holeInRoof \rightarrow wetFloor$.

$$
\begin{aligned}
&A = \{Mary, Bob, Tom\}\\
&\vec{e} \models B_{Mary}wetFloor \wedge B_{Mary}holeInRoof \wedge B_{Mary}rain\\
&\vec{e} \models B_{Mary}B_{Bob}\neg wetFloor \wedge B_{Mary}B_{Bob}\neg rain \wedge B_{Mary}B_{Bob}holeInRoof\\
&\vec{e} \models B_{Mary}B_{Tom}\neg wetFloor \wedge B_{Mary}B_{Tom}rain \wedge B_{Mary}B_{Tom}\neg holeInRoof\\
&\vec{e} \models B_{Mary}\mathit{Expl}(Bob, rain, wetFloor)\\
&\vec{e} \models B_{Mary}\mathit{Expl}(Tom, holeInRoof, wetFloor)
\end{aligned}
$$

We also assume that the agents are able to draw at least simple inferences (and each knows that the others will) and their belief revision operators behave in a sensible way (and each knows that the others' operators do so).

We define a relation $\approx$ that can be understood intuitively as equating two epistemic states, e_i and e_j. For $e_i \approx e_j$ to hold, the internal structures of the states e_i and e_j need not be the same, but they must support the same beliefs as each other, and must continue to do so after any sequence of revisions. Formally, we say that $e_i \approx e_j$ if

- $\vec{e} \models B_i\varphi$ *iff* $\vec{e} \models B_j\varphi$
- and for any sequence of formulas $\alpha_1, \ldots, \alpha_k$, we have that $\vec{e} \models [\alpha_1]_i \cdots [\alpha_k]_i B_i\varphi$ iff $\vec{e} \models [\alpha_1]_j \cdots [\alpha_k]_j B_j\varphi$

Theorem 2. *Given epistemic states $\vec{e}$ and explanandum β, if $e_i \approx e_j$ it then follows that for all α, $\vec{e} \models \mathit{Expl}(i, \alpha, \beta)$ iff $\vec{e} \models \mathit{Expl}(j, \alpha, \beta)$.*

Proof. Note that $\vec{e} \models \mathit{Expl}(i, \alpha, \beta)$ iff $\vec{e} \models [\alpha]_i B_i\beta$ and $\vec{e} \models [\alpha]_i\neg B_i\bot$, and similarly for agent j. The result follows from the definition of $\approx$.

That is, when $e_i \approx e_j$, an objective explanation for the former is also an objective explanation for the latter. Therefore, agent i, acting as the explainer, need not employ its Theory of Mind and reason about agent j's beliefs in order to generate explanations for the latter. However, the fact that $e_i \approx e_j$ does not mean that e_i holds accurate beliefs pertaining to how e_j revises its beliefs. Thus, while any α that explains β may be an objective explanation for both agents i and j, agent i need not necessarily *believe* that α is an explanation for j. Nonetheless, $e_i \approx e_j$ is quite strong, as illustrated by the following theorem.

Theorem 3. *Suppose e_j supports positive and negative introspection – i.e., $\vec{e} \models (B_j\varphi \equiv B_jB_j\varphi) \wedge (\neg B_j\varphi \equiv B_j\neg B_j\varphi)$. Then if $e_i \approx e_j$, agent i will have correct beliefs about j's beliefs, i.e., $\vec{e} \models (B_j\varphi \equiv B_iB_j\varphi) \wedge (\neg B_j\varphi \equiv B_i\neg B_j\varphi)$.*

Proof. If agent j believes φ, then we'll have that $\vec{e} \models B_jB_j\varphi$ (by positive introspection) and then $\vec{e} \models B_iB_j\varphi$ (because $i \approx j$). Similarly, if agent j disbelieves φ, then $\vec{e} \models B_j\neg B_j\varphi$ (by negative introspection) and so $\vec{e} \models B_i\neg B_j\varphi$.

In some cases, an explanation need not cause the explanandum to be entailed by the epistemic state, but rather cause it to be *possible* in the epistemic state. This type of explanation is similar to Boutilier and Becher's *might explanation*.

Definition 5 (Inconsistency-resolving Explanation). *Given epistemic states $\vec{e}$, we say that α explains the possibility of β for agent i if $\vec{e} \models [\alpha]_i\neg B_i\neg\beta$.*

This is a weaker form of explanation but important in various settings such as when an agent is attempting to find an explanation that will allow the behavior of another agent or in consistency-based diagnosis, where the agent is attempting to identify the abnormal components in a system that allow for the observed behavior of the system.

Theorem 4. *Given epistemic states $\vec{e}$ and explanandum β, then for all α, if $\vec{e} \models Expl(i, \alpha, \beta)$ it then follows that α is an inconsistency-resolving explanation for β for agent i, assuming that $\vec{e} \models [\alpha]_i\big((B_i\beta \wedge B_i\neg\beta) \rightarrow B_i\bot\big)$, i.e., that the agent can perform enough reasoning to notice the inconsistency in believing both β and $\neg\beta$.*

This follows straightforwardly from Definitions 3 and 5.

Explanations Involving Agent Beliefs

Importantly, an explainer can utilize its Theory of Mind to generate explanations pertaining to the mental states of other agents, such as their beliefs or goals.

Example 2. Let us reconsider our example where this time, after Mary explains *wetFloor* to Bob, he asks her why Tom doesn't know *wetFloor*. That is, the explanandum β is $\neg B_{Tom}wetFloor$. A possible explanation is then $B_{Tom}\neg holeInRoof$, assuming Bob believes $B_{Tom}rain$.

3.3 Explanations Involving Multiple Agents

An interesting setting that is straightforwardly captured by our framework is one in which an explainer (or explainers) is attempting to explain multiple (possibly disparate) explanandums to multiple explainees.

Definition 6. *Given epistemic states $\vec{e}$ and explanandums β_j, β_k, $\ldots\beta_l$, we say that α explains β_j, β_k, $\ldots\beta_l$ from agent i's perspective for agents j, k, $\ldots l$, respectively, if $\vec{e} \models B_iExpl(j, \alpha, \beta_j) \wedge B_iExpl(k, \alpha, \beta_k) \wedge \ldots \wedge B_iExpl(l, \alpha, \beta_l)$.*

Consider a collaborative card game (e.g., Hanabi [2]) where a certain player is attempting to make different players (each with a unique epistemic state) understand different things with a single piece of information about another player's cards, publicly announced to all players. The explaining player should therefore find an α that explains different explanandums for the different players, given the explaining player's beliefs about the other players' beliefs.

Example 3. In a simpler setting such as our running example, if Mary is trying to explain $wetFloor$ to Bob and Tom at the same time, the explanation α could be $rain \wedge holeInRoof$, where the explanandum for both Bob and Tom is $wetFloor$.

Privacy. Our framework can also capture a notion of privacy. For example, the explainer (agent i) may want to generate an explanation α that explains the explanandum β to some agents (agent j) but not to others (agent k):

$$\vec{e} \models B_i \mathit{Expl}(j, \alpha, \beta) \wedge B_i \neg(\mathit{Expl}(k, \alpha, \beta))$$

Example 4. If Mary, for some reason, wants only Bob to entail $wetFloor$, the explanation α could be $rain$ in which case Bob will entail $wetFloor$ but Tom will not. One can imagine parent #1 wanting to explain something to parent #2 such that their child does not understand.

Multiple Explainers and 'Nested' Explanations. In some cases, there may be multiple explainers trying to explain an explanandum β to an explainee. For example, agents i and j may want to find an α that explains β for agent k:

$$\vec{e} \models B_i \mathit{Expl}(k, \alpha, \beta) \wedge B_j \mathit{Expl}(k, \alpha, \beta)$$

Definition 6 can be straightforwardly extended to capture this setting. Finally, agent i may want to find an α that he believes that agent j believes is an explanation for agent k:

$$\vec{e} \models B_i B_j \mathit{Expl}(k, \alpha, \beta)$$

4 "Best" Explanations for Whom?

An explanadum can typically be explained by a variety of different explanations, but it is often the case that an agent *prefers* one explanation to another relative to some set of criteria. Indeed, there is a large body of previous work (e.g., [4,26,27]) that outlines criteria for defining preference orderings over explanations. In the context of a multiple agents, we have seen that what constitutes an explanation for one agent, may not constitute an explanation for another. This observation extends to the notion of preferred explanations—what's good in the eyes of the explainer may not be good for the explainee, or for all explainees. We explore the issue of preferred explanations briefly here in the context of Theory of Mind.

For each agent in the set of agents A, we define a binary preference relation $\prec$ over explanations such that $\prec_i$ is the preference relation for agent i.

Definition 7 (Preferred Explanation). *Given epistemic states $\vec{e}$ and explanandum β, if α and α' both explain β for agent i and $\alpha \preceq_i \alpha'$, we say that α is at least as preferred as α' for agent i. $\alpha \prec_i \alpha'$ denotes that α is strictly preferred to α' for agent i.*

Similarly, we use $\alpha \preceq_{i,j} \alpha'$ to denote that agent i believes that α is at least as preferred as α' for agent j.

Definition 8 (Optimal Explanation). *Given epistemic states $\vec{e}$ and explanandum β, α is an optimal explanation for β wrt $\prec_i$ iff α explains β for agent i and there does not exist an explanation α' for β for agent i such that α' $\prec_i$ α.*

Hilton [22] posits that an explanation given by one agent to another is a form of conversation and should therefore adhere to Grice's [15] maxims which he proposed as part of a model for effective cooperative conversation. In what follows, we discuss a number of criteria for preferred explanations and relate them to Grice's maxims.

Truthfulness: Grice's first maxim is the **quality** maxim, according to which one must not provide information (e.g., to the explainee) that she believes to be false.

Definition 9 (Subjectively Truthful Explanation). *Given epistemic states $\vec{e}$ and an explanandum β, α is a subjectively truthful explanation for agent j from the perspective of agent i iff $\vec{e} \models B_i Expl(j, \alpha, \beta) \wedge B_i \alpha$.*

Example 5. In our example, Mary may tell Bob that Tom poured water all over the floor, thereby explaining *wetFloor*. However, since Mary does not believe that Tom did such a thing, it would not be a subjectively truthful explanation explanation from Mary's perspective.

Minimality: According to Grice's **quantity** and **relation** maxims, one must provide information that is relevant, sufficiently informative, and no more informative than needed. In a Theory of Mind context, the sufficiency of information is defined relative to the explainer's beliefs about the explainee's epistemic state and the explainer should therefore find the *minimal* explanation relative to the explainee's epistemic state. A large body of work concerned with explanation has discussed a minimality property which an explanation should satisfy. For example, Levesque [26] defines a syntactic simplicity relation between explanations wherein an explanation is *simpler* than another if it contains fewer propositional letters. Minimal explanations in the semantic sense may be defined relative to a set of possible explanations as those that are implied by all other explanations.

Plausibility: Grice's **quality** maxim also dictates that one should not provide information that is not supported by evidence. When applying this maxim to the beliefs of the explainee, an explainer may wish to consider how likely an explanation is from the point of view of the former. For instance, in our example it is more likely that Bob will accept *rain* as an explanation over the highly unlikely explanation according to which Alan Turing came to visit in the middle of the night and accidentally poured water all over the floor. Therefore, the likelihood of an explanation is an important preference criterion when explaining to ourselves and to others. In the quantitative case, Pearl [33] defines a *most probable explanation* while in a qualitative setting the *plausibility* of explanations may be defined where the most plausible explanations are those that require the 'least' change in the explainee's epistemic state (e.g., [4,39]), which could be defined in various ways, including the degree of held beliefs (e.g., [23]).

5 Explainer-Explainee Discrepancies

To this point our account of explanation has assumed the existence of an explanandum, β, that is in need of explanation for a particular agent. However, in the absence of such a prompt, the explainer may use her Theory of Mind to put herself in the explainee's shoes, so to speak, and to identify *discrepancies* between the beliefs of the explainee and those of the explainer, or perhaps in the case of multiple agents, to identify discrepancies between the beliefs of two agents that the explainer can resolve via an explanation. Discrepancies can also arise from inconsistencies between an agent's beliefs and observations in the world. Such discrepancies are common prompts for explanation in the case of diagnosis (e.g., [4,40]).

Definition 10 (Discrepancy). *Given epistemic states $\vec{e}$, β is a discrepancy between e_i and e_j iff $\vec{e} \models B_i\beta \wedge B_j\neg\beta$.*

That is, agent i believes β while agent j believes $\neg\beta$. The beliefs of agents pertaining to discrepancies can also be represented in our framework.

Definition 11 (Subjective Discrepancy). *Given epistemic states $\vec{e}$, β is a discrepancy between e_i and e_j from the perspective of agent i iff $\vec{e} \models B_i(B_i\beta \wedge B_j\neg\beta)$.*

Example 6. In our example, while Mary believes *wetFloor*, she believes that Bob believes that the floor is not wet (i.e., $\vec{e} \models B_{Mary}(B_{Mary}wetFloor \wedge B_{Bob}\neg wetFloor)$. Thus, *wetFloor* is a discrepancy between Bob and Mary's respective epistemic states from Mary's perspective.

Definition 12 (Subjective Discrepancy-resolving Explanation). *Given epistemic states $\vec{e}$ and a discrepancy β between e_i and e_j from the perspective of agent i, we say that α is a discrepancy-resolving explanation for agent j for β from agent i's perspective if $\vec{e} \models B_i[\alpha]_j\neg B_j\neg\beta$.*

Example 7. A discrepancy-resolving explanation for *wetFloor* for Bob from Mary's perspective is *rain*.

Note that Definition 12 appeals to the weaker inconsistency-resolving explanation defined in Definition 5. Thus, the explainer need not find an α that it believes will allow the explainee to entail the discrepancy. Rather, α should resolve the discrepancy by explaining its possibility.

We cast agent i as the explainer and agent j as the explainee, and distinguish between two types of subjective discrepancies: (1) where β is a discrepancy between e_i and e_j from the explainer's perspective; and (2) where β is a discrepancy between e_i and e_j from the explainee's perspective. In (1), as discussed, the explainer (e.g., Mary) may provide a discrepancy-resolving explanation for β (e.g., *rain*). However, for (2), in order to provide such as explanation the explainer must *believe* that the explainee believes that there exists a discrepancy between e_i and e_j. If the explainer's beliefs about the explainee's beliefs are incomplete or incorrect, the former may not recognize that such a discrepancy exists.

Explainer as Mediator. Definition 11 can be straightforwardly generalized to capture a setting where agent i believes that there exists a discrepancy between e_j and e_k:

$$\vec{e} \models B_i(B_j\beta \wedge B_k\neg\beta)$$

Agent i may also believe that agent j believes that α is an explanation for β for agent k, while also believing that α is not in fact a valid explanation for agent k due to the discrepancy between the epistemic states of agents j and k:

$$\vec{e} \models B_i(B_j \mathit{Expl}(k,\alpha,\beta) \wedge \neg \mathit{Expl}(k,\alpha,\beta))$$

Using Definition 6, agent i may explain the discrepancy to agents j and k. Note that the notion of discrepancy discussed here can easily be extended to encode other, possibly richer notions of discrepancy including the degree to which the epistemic states of two agents are discrepant.

6 The (In)Adequacy of the Explainer's Beliefs

The explainer is limited by the accuracy of its beliefs about the explainee's beliefs and reasoning capabilities. Specifically, the explainer's beliefs about the explainee's beliefs and reasoning capabilities must be accurate 'enough' – *adequate* – for the explainer to generate 'good' explanations wrt the explainee.

Definition 13 (Adequacy). *Given epistemic states $\vec{e}$ and explanandum β, we say that agent i's epistemic state e_i is adequate wrt agent j iff for all α, $\vec{e} \models B_i\mathit{Expl}(j,\alpha,\beta)$ iff $\vec{e} \models \mathit{Expl}(j,\alpha,\beta)$.*

That is, if agent i's epistemic state is adequate wrt agent j and β, then it can generate all explanations (for β) for agent j that are also explanations for agent j in its actual epistemic state, e_j.

Theorem 5. *Given epistemic states $\vec{e}$, explanandum β and $\preceq_{i,j}, \preceq_j$, agent i's perspective of agent j's preference relation and agent j's actual preference relation, respectively, if $\preceq_{i,j} = \preceq_j$ and e_i is adequate wrt agent j and β, then for all α, α is an optimal explanation for agent j from agent i's perspective wrt $\preceq_{i,j}$ iff α is an optimal explanation for agent j wrt $\preceq_j$.*

That is, when e_i is adequate wrt agent j and when agent i's beliefs about agent j's preference relation are correct, the optimal explanation for agent j from the perspective of agent i is also the optimal objective explanation for agent j. The proof follows straightforwardly from Definitions 8 and 13.

6.1 Sources of (In)Adequacy

Since most agents do not have a perfect image of another agent's mental state, an agent's beliefs about another agent may be inadequate for a myriad of reasons, including the inaccuracy of an agent's beliefs about the beliefs of other agents and about the way in which other agents revise their beliefs and perform entailment. In what follows, we focus on a setting where an agent holds inadequate beliefs about another agent's beliefs and illustrate using our running example.

Example 8. Returning to our example, assume that Mary forgot that Bob found the hole with her and so she now *falsely* believes that Bob believes that there is no hole in the roof (i.e., $\vec{e} \models B_{Mary}B_{Bob}\neg holeInRoof$). Mary will therefore believe that *rain* ∧ *holeInRoof* is the minimal explanation for Bob (relative to an intuitive measure of minimality). Notice, however, that in her explanation, Mary is conveying more information than is needed for Bob to entail *wetFloor* (thereby violating Grice's quantity maxim).

Example 9. Now consider that Mary *falsely* believes that Bob believes that it had rained and that there is no hole in the roof (perhaps she confused him with Tom!). Mary will therefore believe that *holeInRoof* is an explanation for Bob. However, $\vec{e} \not\models Expl(Bob, holeInRoof, wetFloor)$ since Bob does not believe *rain*. This time, Mary has violated the quantity maxim by not providing *enough* information for Bob to entail *wetFloor*.

Example 10. Mary now falsely believes that Bob believes *wetFloor* (i.e., $\vec{e} \models B_{Mary}B_{Bob}wetFloor$) and so does not provide him with an explanation, believing he does not require one. In this case, while *wetFloor* is an objective discrepancy between Bob and Mary's epistemic states, it is not a discrepancy from Mary's perspective due to her false beliefs.

Addressing Inadequacy
It is possible to mitigate for the inadequacy of the explainer's beliefs in a variety of ways. For example, it may be beneficial for the explainer to attempt to refine its beliefs about the beliefs of the explainee when explanations are not understood by the explainee. To this end, the explainer could try to gather additional pertinent information by acting in the world (e.g., querying the explainee). Additionally, Sreedharan et al. [47] propose a learning technique which enables an explainer to learn a simple model of an explainee and decide, based on the learned model, what information would constitute a good explanation. Further, Sreedharan et al. [46] show how an explainer may generate explanations that are applicable to a set of possible explainee models which arise as the consequence of explainer uncertainty pertaining to the explainee's model.

Finally, while we emphasized the importance of the explainer modelling the beliefs of the explainee, our general account could in theory support the explainee, perhaps compensating for the explainer's inadequate beliefs, reasoning about the beliefs of the explainer to understand a given explanation that might otherwise be construed as inadequate. For example, consider Chandrasekaran et al.'s [9] discussion of a Theory of AI's Mind where a human attempting to better understand a black-box decision making system can do so by familiarizing themselves with the system's capabilities, peculiarities, and shortcomings.

7 Related Work

As previously discussed, we are not the first to propose an account of explanation in terms of the epistemic state of an agent.Levesque presents a knowledge-level account of abduction based on the epistemic state of an agent [26]. He provides a generic definition of explanation that does not commit to a specific type of agent belief. Then, building on his seminal work on a logic of implicit and explicit belief [25], he shows how such different formal models of belief lead to different forms of abductive inference and resultant explanations. Boutilier and Becher [4] similarly appeal to epistemic states to characterize the beliefs of an agent, employing belief revision to allow for explanations that are inconsistent with the epistemic state of the explainee. Prior to the works of Levesque and Boutilier and Becher, Gärdenfors [12] proposed a model of explanation where explanations are defined relative to the epistemic states of agents. While Gärdenfors's account is probabilistic, the models proposed by Levesque and Boutilier and Becher are qualitative. We share the use of epistemic states with all three works, the appeal to qualitative criteria with Levesque and Boutilier and Becher, and the recognition of the importance of belief revision with Boutilier and Becher. Nevertheless, these works all characterize explanation with respect to a single agent providing no account of the distinct beliefs of the explainee *and* explainer, nor do they capture their Theory of Mind.

Nepomuceno-Fernández et al. [32] propose an account of explanation that also recognizes the importance of a revision operator and the use of epistemic states. However, while their Dynamic Epistemic Logic (DEL) based framework can capture multiple agents, their focus remains on an agent's task of obtaining an abductive explanation for itself, rather than for other agents.

Halpern and Pearl [17] proposed a structural model of explanation selection based on the epistemic state of the explainee. In their work, the explainee's epistemic state comprises a set of situations the explainee considers possible and an explanation is then meant to remove some of these possible situations such that the cause of some explanandum may be uniquely identified. Miller extends Halpern and Pearl's approach to include *contrastive* explanations which are given relative to some counterfactual (e.g, in response to the question '*Why P rather than Q?*') [29]. Halpern and Pearl, however, do not discuss some of the necessary elements of Theory of Mind in explanation, such as the notions of explainer-explainee discrepancies and the adequacy of the explainer's beliefs.

In the context of XAIP, Sreedharan et al. [47] demonstrate how the model reconciliation paradigm, proposed by Chakraborti et al. [8], can be generalized to address the important case where the explainee's model of the explainer's planning model is not explicitly known or not provided in a declarative form. Our work captures some of the insights in Sreedharan et al.'s work, in addition to incorporating the notions of epistemic states and belief revision, which in turn allows us to draw inspiration from the rich body of previous work in the field where these ideas originated.

The vast body of work on Theory of Mind proposes two accounts of the way in which agents attribute mental states to other agents: Theory-Theory of Mind [13] (where an agent pre-assigns beliefs to other agents) and Simulation Theory of Mind [14] (where an agent simulates other agents' beliefs and the mechanisms by which those beliefs change). Related work in XAI has highlighted the interesting distinctions between the two as well as the implications for explanation [50]. Further, Sarkadi et al. [42] combine the two approaches by allowing an agent to both assign beliefs to another agent and update its beliefs about the beliefs of the other agent's beliefs by employing Simulation Theory of Mind. We similarly combine the two approaches.

We have focused discussion on the subset of work that is most closely related to the contributions of the paper. For a comprehensive survey of research on explanation, the reader is directed to [30].

8 Concluding Remarks

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 identified a set of desiderata for explanation that utilizes Theory of Mind. These desiderata informed our proposed belief-based account of explanation. Key features of this account are the appeal to epistemic states to capture the mental states of *both* the explainer and explainee, and the use of the explainee's belief revision to assimilate explanations. Further, we formalized and discussed the notion of a discrepancy as a property that allows the explainer to anticipate and provide explanations without prompting. We also presented properties relating to the adequacy of the explainer's beliefs with respect to providing an explanation.

This paper has provided a general characterization of explanation without focusing on its computational realization. This was 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 [26] and Boutilier and Becher [4], our notion of explanation may be realized via an augmentation of existing abductive reasoning systems such as Theorist Poole [35], for example.

Further, in much of this paper we have been relating 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 [31], thereby allowing the explainee to assimiliate this explanation into their beliefs and better interpret the system's decision. As has been argued in this paper, 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 [47] (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.

Acknoweldgements. The authors gratefully acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research (CIFAR), and Microsoft Research.

References

1. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction and revision functions. J. Symb. Logic **50**(2), 510–530 (1985)
2. Bard, N., et al.: The hanabi challenge: a new frontier for AI research. AIJ **280**, 103216 (2020)
3. 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)
4. Boutilier, C., Becher, V.: Abduction as belief revision. AIJ **77**(1), 43–94 (1995)
5. Brachman, R.J., Levesque, H.J.: Knowledge Representation and Reasoning. Elsevier, Amsterdam (2004)
6. Cawsey, A.: Generating interactive explanations. In: AAAI, pp. 86–91 (1991)
7. Chajewska, U., Halpern, J.Y.: Defining explanation in probabilistic systems. arXiv preprint arXiv:1302.1526 (2013)
8. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In: IJCAI, pp. 156–163 (2017)
9. Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., Parikh, D.: It takes two to tango: towards theory of AI's mind. arXiv preprint arXiv:1704.00717 (2017)
10. Charniak, E., McDermott, D.: Introduction to Artificial Intelligence. Addison Wesley, Boston (1985)
11. Darwiche, A., Pearl, J.: On the logic of iterated belief revision. AIJ **89**(1–2), 1–29 (1997)
12. Gärdenfors, P.: Knowledge in Flux: Modeling the Dynamics of Epistemic States. The MIT Press, Cambridge (1988)
13. Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D.: A theory of causal learning in children: causal maps and Bayes nets. Psychol. Rev. **111**(1), 3 (2004)
14. Gordon, R.M.: Folk psychology as simulation. Mind Lang. **1**(2), 158–171 (1986)
15. Grice, H.P.: Logic and conversation. In: Speech Acts, pp. 41–58. Brill (1975)
16. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.: XAI - explainable artificial intelligence. Sci. Robot. **4**(37) (2019)
17. Halpern, J.Y., Pearl, J.: Causes and explanations: a structural-model approach. Part ii: explanations. Br. J. Philos. Sci. **56**(4), 889–911 (2005)
18. Harbers, M., Van den Bosch, K., Meyer, J.J.: Modeling agents with a theory of mind: theory-theory versus simulation theory. Web Intell. Agent Syst. Int. J. **10**(3), 331–343 (2012)
19. Harman, G.H.: The inference to the best explanation. Philos. Rev. **74**(1), 88–95 (1965)
20. Hayes-Roth, F., Waterman, D.A., Lenat, D.B. (eds.): Building Expert Systems. Teknowledge Series in Knowledge Engineering. Addison-Wesley, Boston (1983)
21. Hempel, C.G., Oppenheim, P.: Studies in the logic of explanation. Philos. Sci. **15**(2), 135–175 (1948)

22. Hilton, D.J.: Conversational processes and causal explanation. Psychol. Bull. **107**(1), 65 (1990)
23. van der Hoek, W., Meyer, J.-J.C.: Graded modalities in epistemic logic. In: Nerode, A., Taitslin, M. (eds.) LFCS 1992. LNCS, vol. 620, pp. 503–514. Springer, Heidelberg (1992). https://doi.org/10.1007/BFb0023902
24. 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)
25. Levesque, H.J.: A logic of implicit and explicit belief. In: AAAI, pp. 198–202 (1984)
26. Levesque, H.J.: A knowledge-level account of abduction. In: IJCAI, pp. 1061–1067 (1989)
27. Lipton, P.: Contrastive explanation. Roy. Inst. Philos. Suppl. **27**, 247–266 (1990)
28. McGuinness, D.L., da Silva, P.P.: Explaining answers from the semantic web: the inference web approach. J. Web Semant. **1**(4), 397–413 (2004)
29. Miller, T.: Contrastive explanation: a structural-model approach. arXiv preprint arXiv:1811.03163 (2018)
30. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. AIJ **267**, 1–38 (2019)
31. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Proc. **73**, 1–15 (2018)
32. Nepomuceno-Fernández, A., Soler-Toscano, F., Velázquez-Quesada, F.R.: Abductive reasoning in dynamic epistemic logic. In: Magnani, L., Bertolotti, T. (eds.) Springer Handbook of Model-Based Science. SH, pp. 269–293. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-30526-4_13
33. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Elsevier, Amsterdam (2014)
34. Peirce, C.: Deduction, induction and hypothesis. Pop. Sci. Mon. **13**, 470–482 (1878)
35. Poole, D.: Explanation and prediction: an architecture for default and abductive reasoning. Comput. Intell. **5**(2), 97–110 (1989)
36. Poole, D.: A methodology for using a default and abductive reasoning system. Int. J. Intell. Syst. **5**(5), 521–548 (1990)
37. Pople, H.E.: On the mechanization of abductive logic. In: IJCAI, pp. 147–152 (1973)
38. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behav. Brain Sci. **1**(4), 515–526 (1978)
39. Quine, W.V.O., Ullian, J.S.: The Web of Belief. Random House, New York (1978)
40. Reiter, R.: A theory of diagnosis from first principles. AIJ **32**(1), 57–95 (1987)
41. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)
42. Sarkadi, Ş., Panisson, A.R., Bordini, R.H., McBurney, P., Parsons, S., Chapman, M.: Modelling deception using theory of mind in multi-agent systems. AI Commun. **32**(4), 287–302 (2019)
43. Shortliffe, E.H., Buchanan, B.G.: Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Boston (1985)
44. Slugoski, B.R., Lalljee, M., Lamb, R., Ginsburg, G.P.: Attribution in conversational context: effect of mutual knowledge on explanation-giving. Eur. J. Soc. Psychol. **23**(3), 219–238 (1993)

45. Sohrabi, S., Baier, J.A., McIlraith, S.A.: Preferred explanations: theory and generation via planning. In: AAAI (2011)
46. Sreedharan, S., Chakraborti, T., Kambhampati, S.: Handling model uncertainty and multiplicity in explanations via model reconciliation. In: ICAPS, pp. 518–526 (2018)
47. Sreedharan, S., Hernandez, A.O., Mishra, A.P., Kambhampati, S.: Model-free model reconciliation. In: IJCAI, pp. 587–594 (2019)
48. Stalnaker, R.: The problem of logical omniscience, I. Synthese **89**(3), 425–440 (1991)
49. Weiner, J.: Blah, a system which explains its reasoning. AIJ **15**(1–2), 19–48 (1980)
50. Westberg, M., Zelvelder, A., Najjar, A.: A historical perspective on cognitive science and its influence on XAI research. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) EXTRAAMAS 2019. LNCS (LNAI), vol. 11763, pp. 205–219. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30391-4_12