Tailoring Explanations Through Conversation

Jieting Luo[1], Thomas Studer[2] and Mehdi Dastani[3]

[1] Zhejiang University, Hangzhou, China
luojieting@zju.edu.cn
[2] University of Bern, Bern, Switzerland
thomas.studer@unibe.ch
[3] Utrecht University, Utrecht, the Netherlands
M.M.Dastani@uu.nl

Abstract. The increasing applications of AI systems require personalized explanations for their behaviors to various stakeholders since the stakeholders may have various backgrounds. In general, a conversation between explainers and explainees not only allows explainers to obtain explainees' background, but also allows explainers to tailor their explanations so that explainees can better understand the explanations. In this paper, we propose an approach for an explainer to tailor and communicate personalized explanations to an explainee through having consecutive conversations with the explainee. We prove that the conversation terminates due to the explainee's justification of the initial claim as long as there exists an explanation for the initial claim that the explainee understands and the explainer is aware of.

Keywords: Explanation; Personalization; Conversation; Justification Logic; Modal Logic

1 Introduction

Explainable artificial intelligence (XAI) is one of the important topics in artificial intelligence due to the recognition that it is important for humans to understand the decisions or predictions made by AI systems [19]. Understanding the behavior of AI systems does not only improve the user experience and trust in such systems, but it also allows engineers to better configure them when their behavior needs to change. This is particularly important when AI systems are used in safety-critical domains such as healthcare where decisions made or influenced by clinical decision support systems ultimately affect human life and well-being. In such systems, related stakeholders and professionals must understand how and why certain decisions are made [1].

An important requirement for explainable AI systems is to ensure that the stakeholders with various backgrounds understand the provided explanations, the underlying rationale and inner logic of the decision or prediction process [27]. For example, an explanation of why we should drink enough water throughout the day that is formulated in specialized medical terms may be understandable to medical professionals, but not to young children who have no medical knowledge or background. Therefore, AI systems should be able to provide personalized and relevant explanations that match the backgrounds of their stakeholders. What we mean by personalization is tailoring explanations to accommodate specific users, instead of groups of users. How can the AI systems obtain the backgrounds of their stakeholders? Hilton argues that an explanation is a *social* process of conveying why a claim is made to someone [11]. It is the conversations between explainers and explainees that allow explainers to obtain explainees' background and enable explainers to tailor their explanations so that explainees can better understand the explanations.

In this paper, we propose a novel approach to automatically construct and communicate personalized explanations through the conversation between an explainer and an explainee. Our approach exploits tools and results from justification logic [4,5,15]. The first justification logic, the Logic of Proofs, has been introduced by Artemov to give a classical provability interpretation to intuitionistic logic [2]. Later, various possible worlds semantics for justification logic have been developed [3,8,14,16], which led to epistemic interpretations of justification logic. We will use a logic that features both the modal $\Box$-operator and explicit justification terms t, a combination that goes back to [6]. Our approach is built on the idea of reading *an agent understands an explanation $\mathcal{E}$ for a claim F* as *the agent has a justification t for F such that t also justifies all parts of $\mathcal{E}$*. This is similar to the logic of knowing why where *knowing why F* is related to *having a justification for F* [32]. With this idea, the explainer can interpret the explainee's background from her feedback regarding whether she understands the explanation that has just been received, and provide further explanations given what she has learned about the explainee.

We first develop a two-agent modular model that allows us to represent and reason about agents' knowledge and justification. We then model how the explainee gains more justified knowledge from explanations, and how the explainer specifies her preferences over available explanations using specific criteria. We finally model the conversation where the explainee provides her feedback on the explanation that has just been received and the explainer constructs a further explanation given her current knowledge about the explainee's background interpreted from the explainee's feedback. Our approach ensures that the conversation will terminate due to the explainee's justification of the initial claim as long as there exists an explanation for the initial claim that the explainee understands and the explainer is aware of. This paper is an extended version of [18]. In this version, the definition of explanations is revised to be more general, and the explainer can reason about the explainee's feedback to gain information about the explainee's background according to our definition of understanding explanations, while in [18] the explainer does not perform the reasoning but directly gains the information about the explainee's background, which is less intuitive. Furthermore, we have defined two operations between explanations, enabling us to characterize the criteria that we design for explanation evaluation.

2 Multi-agent Modular Models

Let Prop be a countable set of atomic propositions. The set of propositional formulas $\mathcal{L}_{\mathrm{Prop}}$ is inductively defined as usual from Prop, the constant $\bot$, and the binary connective $\to$. We now specify how we represent justifications and what operations on them we consider. We assume a countable set $\mathrm{JConst} = \{c_0, c_1, \ldots\}$ of justification constants. Further, we assume a countable set JVar of justification variables, where each variable is indexed by a propositional formula and a (possibly empty) list of propositional formulas, i.e. if $A_1, \ldots, A_n, B \in \mathcal{L}_{\mathrm{Prop}}$, then $x_B^{A_1,\ldots,A_n}$ is a justification variable. Constants denote atomic justifications that the system no longer analyzes, and variables denote unknown justifications. Justification terms are defined inductively as follows:

$$t ::= c \mid x \mid t \cdot t$$

where $c \in \mathrm{JConst}$ and $x \in \mathrm{JVar}$. We denote the set of all terms by Tm. A term is ground if it does not contain variables, so we denote the set of all ground terms by Gt. A term can be understood as a proof or an evidence. Let Agt be a finite set of agents. Formulas of the language $\mathcal{L}_{\mathrm{J}}$ are defined inductively as follows:

$$A ::= p \mid \bot \mid A \to A \mid \Box_i A \mid [\![t]\!]_i A$$

where $p \in \mathrm{Prop}$, $i \in Agt$ and $t \in \mathrm{Tm}$. Formula $\Box_i A$ is interpreted as "agent i knows A", and formula $[\![t]\!]_i A$ is interpreted as "agent i uses t to justify A".

The model we use in this paper is a two-agent modular model that interprets justification logic in a two-agent context. It is a Kripke model extended with two agents and two evidence functions. In general, evidence accepted by different agents is distinct, so evidence terms for each agent are constructed using her own basic evidence function.

Definition 2.1 (Two-agent Modular Models). *A two-agent modular model over a set of atomic propositions* Prop *and a set of terms* Tm *is defined as a tuple* $\mathcal{M} = (Agt, W, \tilde{R}, \tilde{*}, \pi)$, *where*

- $Agt = \{1, 2\}$ *is a set of agents, and we assume that it is always the case that agent 1 announces an explanation to agent 2;*
- $W \neq \varnothing$ *is a set of worlds;*
- $\tilde{R} = \{R_1, R_2\}$ *for each agent in Agt, where* $R_i \subseteq W \times W$ *is a reflexive and transitive accessibility relation;*
- $\tilde{*} = \{*_1, *_2\}$ *for each agent in Agt, where* $*_i : \mathrm{Tm} \times W \to \mathcal{P}(\mathcal{L}_{\mathrm{Prop}})$ *is an evidence function that maps a term* $t \in \mathrm{Tm}$ *and a world* $w \in W$ *to a set of formulas in* $\mathcal{L}_{\mathrm{Prop}}$*;*
- $\pi : \mathrm{Prop} \to 2^W$ *is an evaluation function for the interpretation of propositions.*

We assume that the agents have finite reasoning power. Therefore, we restrict our models such that for each $w \in W$ and agent i,

- there are only finitely many $t \in \mathrm{Gt}$ such that $*_i(t, w)$ is non-empty, and
- for each $t \in \mathrm{Gt}$, the set $*_i(t, w)$ is finite.

Moreover, it is not necessary that

$$(F \to G) \in *_i(s, w) \text{ and } F \in *_i(t, w) \text{ imply } G \in *_i(s \cdot t, w) \qquad (\dagger)$$

Definition 2.2 (Truth Evaluation). *We define what it means for a formula* A *to hold under a two-agent modular model* $\mathcal{M}$ *and a world* w*, written as* $\mathcal{M}, w \vDash A$*, inductively as follows:*

- $\mathcal{M}, w \nvDash \bot$*;*

- *$\mathcal{M}, w \models P$ iff $w \in \pi(P)$;*
- *$\mathcal{M}, w \models F \rightarrow G$ iff $\mathcal{M}, w \not\models F$ or $\mathcal{M}, w \models G$;*
- *$\mathcal{M}, w \models \Box_i F$ iff for any $u \in W$, if wR_iu, then $\mathcal{M}, u \models F$;*
- *$\mathcal{M}, w \models [\![t]\!]_i F$ iff $F \in *_i(t, w)$.*

Other classical logic connectives (e.g.,"$\wedge$", "$\vee$") are assumed to be defined as abbreviations by using $\perp$ and $\rightarrow$ in the conventional manner. We say that a formula A is valid in a model $\mathcal{M}$, written as $\mathcal{M} \models A$ if $\mathcal{M}, w \models A$ for all $w \in W$. We say that a formula A is a validity, written as $\models A$ if $\mathcal{M} \models A$ for all models $\mathcal{M}$.

We require that modular models satisfy the property of *justification yields knowledge*: for any ground term t, agent i, and world w, if $F \in *_i(t, w)$, then for any $u \in W$, if wR_iu, then $\mathcal{M}, u \models F$, which gives rise to the following validity:

$$\models [\![t]\!]_i F \rightarrow \Box_i F. \tag{JYK}$$

Note that in contrast to usual models of justification logic, we require justification yields knowledge only for ground terms (and not for all terms as is originally required in modular models). The reason is that we interpret justification variables in a new way. Traditionally, a justification variable stands for an arbitrary justification. Hence $[\![x]\!]_i F \rightarrow \Box_i F$ should hold: no matter which justification we have for F, it should yield knowledge of F. In this paper, we use a different reading of justification variables. They stand for open assumptions, which do not (yet) have a justification. Therefore, $[\![x]\!]_i F$ will not imply knowledge of F. Our modular model gives rise to the following validity due to the reflexivity of the accessibility relations:

$$\models \Box_i F \rightarrow F.$$

Combining this with validity JYK, we find that justifications by ground terms are factive: for any ground term t and any formula F, we have

$$\models [\![t]\!]_i F \rightarrow F.$$

Notice that our model does not respect the usual application operation (·) on evidence terms due to the removal of constraint (†) from our model,

$$\not\models [\![s]\!]_i (F \rightarrow G) \rightarrow ([\![t]\!]_i F \rightarrow [\![s \cdot t]\!]_i G).$$

This is because agents are limited in their reasoning powers and thus might not be able to derive all of the logical consequences of their justified knowledge by constructing proofs, which is why agents need explanations.

3 Understanding and Learning from Explanations

Given a claim, an agent can construct a deduction for a claim, which we call an explanation of the claim in this paper. An explanation is inductively defined as a tree of formulas.

Definition 3.1 (Explanations). *Explanations are inductively defined. An explanation is given by*

$$\frac{\Lambda_1, \ldots, \Lambda_n}{B}$$

where each Λ_i is an explanation or a formula of $\mathcal{L}_{\mathrm{Prop}}$ and B is a formula of $\mathcal{L}_{\mathrm{Prop}}$. Hence, an explanation is a finite tree of formulas with at least two nodes.

Let $\mathcal{E}$ be an explanation. We say that

- *The claim of $\mathcal{E}$, denoted by $claim(\mathcal{E})$, is the root of $\mathcal{E}$;*
- *$Pr(\mathcal{E}, B)$ is the list of premises of B, which is given as the list of the child nodes of B (strictly speaking, we need an occurrence of B in $\mathcal{E}$);*
- *a formula is a hypothesis in $\mathcal{E}$ if it is a leaf node of $\mathcal{E}$, and we use $H(\mathcal{E})$ to denote the set of hypotheses of $\mathcal{E}$;*
- *a formula in $\mathcal{E}$ is called a derived formula if it is not a leaf node of $\mathcal{E}$, and we use $D(\mathcal{E})$ to denote the set of derived formulas of $\mathcal{E}$.*

Having defined an explanation as a tree of formulas, we can define a sub-explanation of a given explanation as a sub-tree.

Definition 3.2 (Sub-explanations). *Given two explanations $\mathcal{E}$ and $\mathcal{E}'$, we say that $\mathcal{E}'$ is a sub-explanation of $\mathcal{E}$ if $\mathcal{E}'$ is a sub-tree of $\mathcal{E}$. We use $sub(\mathcal{E}, F)$ to denote the sub-explanation of $\mathcal{E}$ with $claim(\mathcal{E}') = F$.*

One important property of justification logic is its ability to internalize its own notion of proof. If B is derivable from $A_1, \ldots, A_n$, then there exists a term $t \cdot x_1 \cdot \cdots \cdot x_n$ such that $[\![t \cdot x_1 \cdot \cdots \cdot x_n]\!]_i B$ is derivable from $[\![x_1]\!]_i A_1, \ldots, [\![x_n]\!]_i A_n$.[4] The justification term $t \cdot x_1 \cdot \cdots \cdot x_n$ justifying B represents a blueprint of the derivation of B from $A_1, \ldots, A_n$. In this section, we will define a procedure that mimics the application operation on terms to construct derived terms in order to internalize the deduction of an explanation. Given an explanation, a derived term of the conclusion with respect to the explanation is constructed with the justifications of the premises and the deduction from the premises to the conclusion. Typically, if there exists any premises that the agent cannot justify, then a variable is used for its justification in the derived term; if the agent cannot justify the deduction, then a variable is used as the derived term.

Definition 3.3 (Construction of Derived Terms). *Given a two-agent modular model $\mathcal{M}$, a world w, an explanation $\mathcal{E}$, and a derived formula B occurring in $\mathcal{E}$, we define the explainee's derived term of B with respect to $\mathcal{E}$ inductively as follows:*

- *Case: B is the claim of a simple explanation $\mathcal{E}' = A_1, \ldots, A_n / B$. We distinguish two cases:*
  1. *If there exists $d \in \mathrm{Gt}$ such that $\mathcal{M}, w \vDash [\![d]\!]_2 (A_1 \to (\cdots \to (A_n \to B) \cdots))$, then the derived term of B has the form $d \cdot t_1 \cdots t_n$ where the terms t_i are given by: if there exists $s_i \in \mathrm{Gt}$ with $\mathcal{M}, w \vDash [\![s_i]\!]_2 A_i$, then set $t_i = s_i$; else, set $t_i = x_{A_i}$;*
  2. *otherwise, the derived term of B has the form $x_B^{Pr(\mathcal{E},B)}$.*

[4] This property requires a so-called axiomatically appropriate constant specification.

- *Case: B is the claim of an explanation $\mathcal{E}' = \mathcal{E}'_1, \dots \mathcal{E}'_n / B$. We distinguish two cases:*
  1. *If there exists $d \in \mathrm{Gt}$ such that*
  $$\mathcal{M}, w \vDash [\![d]\!]_2(claim(\mathcal{E}'_1) \to (\cdots \to (claim(\mathcal{E}'_n) \to B)\cdots)),$$
  then the derived term of B has the form $d \cdot t_1 \cdots t_n$ where each t_i is the derived term of $claim(\mathcal{E}'_i)$ with respect to $\mathcal{E}$;
  2. *otherwise, the derived term of B has the form $x_B^{Pr(\mathcal{E},B)}$.*

Example 1. Assume that we have a two-agent modular model $\mathcal{M}$. Agent 2 hears an example $\mathcal{E} = A/B/C$ in world w, and it is the case that

$$\begin{aligned} &\mathcal{M}, w \vDash [\![t_A]\!]_2 A \\ &\mathcal{M}, w \vDash [\![d_{A\to B}]\!]_2 (A \to B) \\ &\mathcal{M}, w \vDash [\![d_{B\to C}]\!]_2 (B \to C) \end{aligned}$$

for ground terms t_A, $d_{A\to B}$, and $d_{B\to C}$, then the derived term of B with respect to $\mathcal{E}$ is $d_{A\to B} \cdot t_A$, and the derived term of C with respect to $\mathcal{E}$ is $d_{B\to C} \cdot (d_{A\to B} \cdot t_A)$. If the explainee cannot justify A, then the derived term of C with respect to $\mathcal{E}$ would become $d_{B\to C} \cdot (d_{A\to B} \cdot x_A)$; if the explainee cannot justify $A \to B$, then the derived term of C with respect to $\mathcal{E}$ would become $d_{B\to C} \cdot x_B^A$.

The explainee's justification for a deduction can be seen as her reasoning capability and can be different from agent to agent. If the explainee cannot justify a deduction step, then the deduction is beyond her reasoning capability. In real life, agents' reasoning capabilities can be limited by factors such as age, profession and experience. For example, a mathematician can follow complicated mathematical proofs, while a primary student can only follow simple mathematical proofs. Further, for a derived formula in an explanation, the explainee might have another term that has nothing to do with the explanation to justify it. But using this term to justify the formula does not mean that the explainee can follow the explanation, so we need to require that a derived term be formed by justification terms that are used to justify its premises and deduction in the explanation. We should also notice that a derived term of a derived formula with respect to an explanation might not be unique, because there might exist multiple terms for the explainee to justify the hypotheses in the explanation, making the derived terms different.

Once an agent hears an explanation, she can update her justification with derived terms that she constructs, which means that the agent learns from the explanation and has more justified knowledge.

Definition 3.4 (Learning from Explanations). *Given a two-agent modular model $\mathcal{M}$, a world w and an explanation $\mathcal{E}$, after the explainee hears $\mathcal{E}$ in w, $\mathcal{M}$ is updated as $\mathcal{M}|(2, w, \mathcal{E})$, where $\mathcal{M}|(2, w, \mathcal{E}) = (Agt, W, \tilde{R}, \tilde{*}', \pi)$ is defined as follows:*

- *for any $F \in D(\mathcal{E})$, $*'_2(t, w) = *_2(t, w) \cup \{F\}$, where t is the explainee's derived term of F with respect to $\mathcal{E}$;*

- *for any $s \in \mathrm{Tm}$ and any $G \in \mathcal{L}_{\mathrm{Prop}}$, if $G \in *_2(s, w)$, then $G \in *'_2(s[r/x_{claim(\mathcal{E})}], w)$ and $G \in *'_2(s[r/x^{H(\mathcal{E})}_{claim(\mathcal{E})}], w)$, where r is the explainee's derived term of $claim(\mathcal{E})$ with respect to $\mathcal{E}$;*
- *for the explainer, $*'_1(\cdot) = *_1(\cdot)$.*

In words, after the explainee hears explanation $\mathcal{E}$ in world w, for each derived formula F in $\mathcal{E}$, the explainee's justification of F will be updated with its derived term t; the derived term r of $claim(\mathcal{E})$ with respect to $\mathcal{E}$ is substituted for every occurrence of $x_{claim(\mathcal{E})}$ and $x^{H(\mathcal{E})}_{claim(\mathcal{E})}$ in the explainee's justification. Recall that agents in this paper have limited reasoning powers and thus might be unable to derive all of the logical consequences from their knowledge. Hearing an explanation allows the explainee to gain new justified knowledge by connecting existing justified knowledge. Note that we do not need to remove the explainee's epistemic access R_2 to worlds where F does not hold to guarantee validity JYK. The reason is as follows: if t is not a ground term, then the explainee has yet to justify F, and thus the explainee's knowledge should remain the same as before learning from explanation $\mathcal{E}$; if t is a ground term, then the explainee knows F due to validity JYK, and the knowledge of F is ensured by the validity of the modal k-axiom (our model still respects the epistemic closure). Other agents' justification and epistemic access remain the same. Apart from gaining new justified knowledge for derived formulas, the explainee can also learn deduction from explanation. For example, the explainee can learn $A \to C$ from $A \to B$ and $B \to C$ in an explanation, gaining more reasoning capability. However, we notice that this is beyond the scope of this paper, and future work on learning from explanations can have more sophisticated formalism on this part.

The learning process gives rise to some intuitive consequences. First of all, when the explainee is already aware of the explanation before it is announced, the explainee will not gain any new justified from the explanation. Secondly, it is possible for the explainee to gain justification for the formulas that are not contained in the explanation, because the learning process contains substituting derived terms of formulas for their corresponding variables, which means that the explainee can justify more than what an explanation has. As standard, the resulting updated model is still a two-agent modular model.

Proposition 3.1. *Given a two-agent modular model $\mathcal{M}$, a world w and an explanation $\mathcal{E}$, after the explainee hears explanation $\mathcal{E}$ in world w, $\mathcal{M}$ is updated as $\mathcal{M}|(2, w, \mathcal{E})$, which is still a two-agent modular model.*

Proof. The update defined in Definition 3.4 only affects the explainee's evidence function, so we need to check whether the updated model still has validity JYK. Suppose the derived terms t and s that function $*_2$ is updated with are ground terms, which means that the explainee can justify all the hypotheses and deduction steps for F and G (as Proposition 3.2 will prove). Since the original model has validity JYK, the explainee knows the hypotheses and deduction steps. Due to the epistemic closure, the explainee also has knowledge of F and G.

We then extend our language $\mathcal{L}_{\mathrm{J}}$ with new formulas of the form $[i : \mathcal{E}]\varphi$, read as "φ is true after agent i hears explanation $\mathcal{E}$", and its evaluation is defined with respect to a two-agent modular model $\mathcal{M}$ and a world w as follows:

$$\mathcal{M}, w \models [i : \mathcal{E}]\varphi \quad \text{iff} \quad \mathcal{M}|(i, w, \mathcal{E}), w \models \varphi.$$

Intuitively, an agent understands an explanation if the derived term of its conclusion does not contain any variables (unknown justification), i.e., it is a ground term.

Definition 3.5 (Understanding Explanations). *Given a two-agent modular model $\mathcal{M}$, a world w and an explanation $\mathcal{E}$, let t be the explainee's derived term of $claim(\mathcal{E})$ with respect to $\mathcal{E}$ in world w. We say that the explainee understands $\mathcal{E}$ in world w iff t is a ground term.*

Example 2. In Example 1, the explainee understands explanation $\mathcal{E}$ if the derived term of C with respect to $\mathcal{E}$ is $d_{B \to C} \cdot (d_{A \to B} \cdot t_A)$; the explainee cannot understand explanation $\mathcal{E}$ if the derived term of C with respect to $\mathcal{E}$ is $d_{B \to C} \cdot (d_{A \to B} \cdot x_A)$ or $d_{B \to C} \cdot x_B^A$.

Given that an explanation is defined as a deduction tree, we have sufficient and necessary conditions for understanding an explanation: an agent understands an explanation if and only if the agent can justify all the hypotheses and deduction steps that are used in the explanation. Let G be a formula and $L = A_1, \ldots, A_n$ be a list of formulas. Then the expression $\underset{L}{\Longrightarrow} G$ stands for $A_1 \to (\cdots \to (A_n \to G)\cdots)$.

Proposition 3.2. *Given a two-agent modular model $\mathcal{M}$, a world w and an explanation $\mathcal{E}$, the explainee understands explanation $\mathcal{E}$ in world w iff*

- *for any $F \in H(\mathcal{E})$, there exists a ground term $t \in \mathrm{Gt}$ such that $\mathcal{M}, w \models [\![t]\!]_2 F$,*
- *for any $G \in D(\mathcal{E})$, there exists a ground term $s \in \mathrm{Gt}$ such that*

$$\mathcal{M}, w \models [\![s]\!]_2 (\underset{Pr(\mathcal{E}, G)}{\Longrightarrow} G).$$

Proof. The explainee understands $\mathcal{E}$ if and only if the derived term of $claim(\mathcal{E})$ that the explainee constructs with respect to $\mathcal{E}$ is a ground term. Given the way of constructing derived terms, the derived term of $claim(\mathcal{E})$ is ground if and only if the explainee can justify all parts of $claim(\mathcal{E})$, i.e., all the hypotheses and deduction steps in $claim(\mathcal{E})$.

Conversely, the agent cannot understand the announced explanation if and only if there exists a hypothesis that the agent cannot justify, or there exists a deduction step that is beyond the agent's reasoning capability, making that the agent cannot construct a ground derived term for each derived formula in the explanation after hearing the explanation. Moreover, if the agent understands an explanation, she is able to justify every derived formula in the explanation with a ground term after hearing the explanation.

Proposition 3.3. *Given a two-agent modular model $\mathcal{M}$, a world w and an explanation $\mathcal{E}$, if the explainee understands explanation $\mathcal{E}$, then for any derived formula $F \in D(\mathcal{E})$, there exists a ground term t such that*

$$\mathcal{M}, w \models [2 : \mathcal{E}][\![t]\!]_2 F.$$

Proof. If the explainee understands $\mathcal{E}$, then the explainee can justify all the hypotheses and deduction steps in $claim(\mathcal{E})$. Given the way of constructing derived terms and learning from explanations, the explainee can then justify any derived formula in $\mathcal{E}$ with a ground term.

However, the reverse direction does not hold, namely having this formula does not necessarily imply that the explainee understands explanation $\mathcal{E}$, because the ground term t might have nothing to do with the explanation. Another important property about learning from an explanation is that for an agent's justified knowledge that have nothing to do with the explanation, it will remain the same after hearing an explanation, and it was also the case before hearing the explanation.

Proposition 3.4. *Given a two-agent modular model $\mathcal{M}$, a world w and an explanation $\mathcal{E}$, if the explainee constructs a derived term t_F for each derived formula F with respect to $\mathcal{E}$, then for any ground term $s \in \mathrm{Gt}$ that does not contain t_F and any formula $P \in \mathcal{L}_{\mathrm{Prop}}$,*

$$\mathcal{M}, w \vDash [\![s]\!]_2 P \leftrightarrow [2:\mathcal{E}][\![s]\!]_2 P.$$

Proof. Direction →: The update approach in Definition 3.4 contains expanding justified knowledge and replacing variable terms with terms that have just been constructed. Since s is a ground term, it will not get replaced after $\mathcal{E}$ is announced. Thus, if the explainee justifies P with s before $\mathcal{E}$ is announced, it is still the case after $\mathcal{E}$ is announced. Namely, $\mathcal{M}, w \vDash [\![s]\!]_2 P \rightarrow [2:\mathcal{E}][\![s]\!]_2 P$. Direction ←: As s does not contain t_F, it is not constructed due to $\mathcal{E}$. Thus, if the explainee justifies P with s after $\mathcal{E}$ is announced, it is already the case before $\mathcal{E}$ is announced. Namely, $\mathcal{M}, w \vDash [\![s]\!]_2 P \leftarrow [2:\mathcal{E}][\![s]\!]_2 P$.

The direction from left to right is the persistence principle saying that an agent's knowledge always get expanded after hearing an explanation. The direction from right to left states that if after hearing an explanation the agent knows P for a reason that is independent on the explanation, then before hearing the explanation the agent already knowd P for the same reason. In other words, having terms in our logical language also allows the agent to distinguish the justified knowledge due to the explanation from the justified knowledge due to another, unrelated reasons, which purely modal logic cannot formulate.

Example 3. Continued with Example 1, suppose the explainee uses term $d_{B\rightarrow C} \cdot x_B^A$ to justify C because she cannot justify the deduction from A to B. After hearing explanation $\mathcal{E}$, it is the case that

$$\mathcal{M}, w \vDash [2:\mathcal{E}][\![d_{B\rightarrow C} \cdot x_B^A]\!]_2 C.$$

The explainee continues to hear another explanation $\mathcal{E}' = A/D/B$, and it is the case that

$$\begin{aligned}
&\mathcal{M}, w \vDash [2:\mathcal{E}][\![t_A]\!]_2 A,\\
&\mathcal{M}, w \vDash [2:\mathcal{E}][\![d_{A\rightarrow D}]\!]_2 (A \rightarrow D),\\
&\mathcal{M}, w \vDash [2:\mathcal{E}][\![d_{D\rightarrow B}]\!]_2 (D \rightarrow B).
\end{aligned}$$

The explainee then can construct the derived term of B with respect to $\mathcal{E}'$ as $d_{D\to B} \cdot (d_{A\to D} \cdot t_A)$,

$$\mathcal{M}, w \vDash [2:\mathcal{E}'][2:\mathcal{E}][\![d_{D\to B} \cdot (d_{A\to D} \cdot t_A)]\!]_2 B.$$

Moreover, according to the learning approach in Definition 3.4, $d_{D\to B} \cdot (d_{A\to D} \cdot t_A)$ is substituted for every occurrence of x^A_B in the explainee's justification. Thus, the derived term of C with respect to $\mathcal{E}$ becomes $d_{B\to C} \cdot (d_{D\to B} \cdot (d_{A\to D} \cdot t_A))$,

$$\mathcal{M}, w \vDash [2:\mathcal{E}'][2:\mathcal{E}][\![d_{B\to C} \cdot (d_{D\to B} \cdot (d_{A\to D} \cdot t_A))]\!]_2 C.$$

Now we can summarize that a user-agent profile consists of her justified knowledge and deductions. It is important for the explainer to gain information about these two aspects of the explainee through having consecutive conversations with the explainee in order to provide an explanation that can be understood by the explainee.

4 Explanation Evaluation

In the previous section, we presented how the explainee's mental state is updated after hearing an explanation. In this section, we will investigate how the explainer selects an explanation for announcing to the explainee. First of all, an explanation is selected from the explanations that the explainer is aware of, namely the explanations where all the hypotheses as well as all the derived formulas are justified by the explainer with ground derived terms with respect to the explanation. Secondly, an explanation that contains information that the explainee cannot justify should not be selected. In order to express these two requirements, we need to extend our language $\mathcal{L}_J$ with new formulas of the form $\triangle_i P$, read as "agent i can justify P", and its evaluation is defined with respect to a two-agent modular model $\mathcal{M}$ and a world w as follows:

- $\mathcal{M}, w \vDash \triangle_i P$ iff there exists $t \in \mathrm{Gt}$ such that $\mathcal{M}, w \vDash [\![t]\!]_i P$.

Compared with formula $[\![t]\!]_i P$, the term t is omitted in formula $\triangle_i P$, meaning that agent i can justify A but we don't care how she justifies P. Because of validity JYK, we have the following validity:

$$\vDash \triangle_i P \to \square_i P.$$

Definition 4.1 (Available Explanations). *Given a two-agent modular model $\mathcal{M}$ and a world w, we say that an explanation $\mathcal{E}$ is available for the explainer to the explainee in world w iff*

- *there exists $t \in \mathrm{Gt}$ such that t is a derived term of $claim(\mathcal{E})$ with respect to $\mathcal{E}$ and $\mathcal{M}, w \vDash [\![t]\!]_1 claim(\mathcal{E})$;*
- *there does not exist $P \in H(\mathcal{E})$ such that $\mathcal{M}, w \vDash \square_1 \neg \triangle_2 P$;*
- *there does not exist $Q \in D(\mathcal{E})$ such that $\mathcal{M}, w \vDash \square_1 \neg \triangle_2 (\underset{Pr(\mathcal{E},Q)}{\Longrightarrow} Q)$.*

Given a formula F as a claim and a set of formulas A as hypotheses, the set of agent 1's available explanations to the explainee for proving F from A is denoted as

$$\lambda_{1,2}^{\mathcal{M},w}(A,F) = \{\mathcal{E} \mid claim(\mathcal{E}) = F, H(\mathcal{E}) = A \text{ if } A \neq \varnothing, \text{ and } \\ \mathcal{E} \text{ is available for the explainer to the explainee given } \mathcal{M} \text{ and } w\}.$$

We write $\lambda_{1,2}^{\mathcal{M},w}(A,F)$ as $\lambda^{\mathcal{M},w}(A,F)$ when the agents are clear from the context.

Compared with the explainee that needs to construct derived terms of derived formulas in an explanation, the explainer has already justified all the derived formulas in an explanation that is available to him by derived terms and makes sure that there does not exist a hypothesis or a deduction that the explainee cannot justify.

If an explanation is available to the explainer, she can announce it to the explainee. But when there are multiple available explanations, the explainer must select one among them given what she knows about the explainee. The question is what criteria the explainer can hold for explanation selection. Notice that the explainer and the explainee are engaged in a conversation, so they are expected to adhere to basic conversational rules. Jaspars and Hilton argue that a good explanation must align with the mental model of the explainee [12][13]. Therefore, the explainer should select an explanation that is most likely to be understood by the explainee. Looking back at our definition of understanding an explanation, we can say that one explanation is more likely to be understood by an agent than another explanation if the former one contains more hypotheses and deductions that are justified by the agent than the latter one. However, this criterion alone is insufficient. Simpler and more general explanations are tested to be better explanations [26][22], and providing more information than necessary would increase the cognitive load of the explainee. Thus, agent 1 is supposed to make short explanations, favoring those with fewer deductive steps (i.e., derived formulas). Since the goal of the explainer is to announce an explanation that can be understood by the explainee, it makes sense for the first criterion to have priority over the second one. In this paper, we impose a total pre-order $\precsim$ over available explanations to represent the preference between two explanations with respect to these two criteria. For better expression, we use $N_{1,2}^{\mathcal{M},w}(\mathcal{E})$ to denote the set of hypotheses and deductions in explanation $\mathcal{E}$ that the explainer is not sure whether the explainee can justify or not.

$$N_{1,2}^{\mathcal{M},w}(\mathcal{E}) = \{F \mid \mathcal{M}, w \vDash \neg \Box_1 \triangle_2 F, \text{ where } F \text{ is a hypothesis in } \mathcal{E}, \text{ or} \\ \mathcal{M}, w \vDash \neg \Box_1 \triangle_2 (\underset{Pr(\mathcal{E},F)}{\Longrightarrow} F), \text{ where } F \text{ is a derived formula in } \mathcal{E}.\}$$

We might write $N^{\mathcal{M},w}(\mathcal{E})$ for short if the agents are clear from the context.

Definition 4.2 (Preferences over Explanations). *Given a two-agent modular model $\mathcal{M}$ and a world w, for any two explanations $\mathcal{E}, \mathcal{E}'$, $\mathcal{E} \precsim^{\mathcal{M},w} \mathcal{E}'$ iff*

- $|N^{\mathcal{M},w}(\mathcal{E})| > |N^{\mathcal{M},w}(\mathcal{E}')|$*; or*
- $|N^{\mathcal{M},w}(\mathcal{E})| = |N^{\mathcal{M},w}(\mathcal{E}')|$ *and* $D(\mathcal{E}) \geq D(\mathcal{E}')$.

As is standard, we also define $\mathcal{E} \sim^{\mathcal{M},w} \mathcal{E}'$ to mean $\mathcal{E} \precsim^{\mathcal{M},w} \mathcal{E}'$ and $\mathcal{E}' \precsim^{\mathcal{M},w} \mathcal{E}$, and $\mathcal{E} \prec^{\mathcal{M},w} \mathcal{E}'$ to mean $\mathcal{E} \precsim^{\mathcal{M},w} \mathcal{E}'$ and $\mathcal{E} \not\prec^{\mathcal{M},w} \mathcal{E}'$. The above definition of preferences between two explanations specifies how the explainer selects an explanation to announce to the explainee: given two available explanations $\mathcal{E}$ and $\mathcal{E}'$, the explainer first compares two explanations in terms of the number of hypotheses and deductions that might not be justified by the explainee, and the one with less number is more preferable; if both explanations have the same number of hypotheses and deductions that might not be justified by the explainee, then the explainer compares these two explanations in terms of the number of deduction steps in the explanations, and the one with less number is more preferable. Using this approach, the explainer always cuts out the part that she knows for sure that agent 2 can justify, or replaces some part of the explanation with a shorter deduction that she knows that the explainee can justify, making explanations shorter.

In order to formalize these properties, we define an operation of replacement between explanations. The idea is that we can replace some part of an explanation $\mathcal{E}$, which can be a sub-explanation or a hypothesis, with another explanation $\mathcal{E}'$ if they claim for the same thing, denoted as $\mathcal{E}[\mathcal{E}']$.

Definition 4.3 (Replacement). *Let $\mathcal{E}'$ be an explanation.*

- *If $F = claim(\mathcal{E}')$, then $F[\mathcal{E}'] := \mathcal{E}'$; otherwise, $F[\mathcal{E}'] := F$.*
- *If $claim(\mathcal{E}') = B$, then*

$$\frac{\Lambda_1,\ldots,\Lambda_n}{B}[\mathcal{E}'] = \mathcal{E}'$$

otherwise,

$$\frac{\Lambda_1,\ldots,\Lambda_n}{B}[\mathcal{E}'] = \frac{\Lambda_1[\mathcal{E}'],\ldots,\Lambda_n[\mathcal{E}']}{B}$$

The operation is defined inductively. If replacement is operated on formula F and explanation $\mathcal{E}'$, then we need to check whether F is the claim of $\mathcal{E}'$. If yes, F is replaced with $\mathcal{E}'$; if no, then F stays there. If explanation $\mathcal{E}'$ has B as its claim, then the whole explanation for B is replaced with $\mathcal{E}'$; otherwise, the replacement is operated on Λ_1 to Λ_n. In the case where F does not occur anywhere in $\mathcal{E}$, the operation returns $\mathcal{E}$. Notice that, if what $\mathcal{E}'$ replaces is a hypothesis in $\mathcal{E}$, the operation can be seen as combining $\mathcal{E}$ and $\mathcal{E}'$. We can see how the operation of replacement is performed from Fig. 1.

Given an explanation $\mathcal{E}$ and a derived formula F in $\mathcal{E}$, if the explainer knows for sure that the explainee can justify F, then the explainer can cut out the deduction for F from $\mathcal{E}$, resulting a shorter explanation $\mathcal{E}'$.

Proposition 4.1. *Given a two-agent modular model $\mathcal{M}$ and a world w, for any two explanations $\mathcal{E}$ and $\mathcal{E}'$, $\mathcal{E} \prec^{\mathcal{M},w} \mathcal{E}'$ if*

- *there exists a formula $F \in D(\mathcal{E})\backslash\{claim(\mathcal{E})\}$ such that $\mathcal{M}, w \vDash \Box_1 \triangle_2 F$,*
- *$\mathcal{E} = \mathcal{E}'[sub(\mathcal{E}, F)]$.*

Proof. Given two explanations $\mathcal{E}$ and $\mathcal{E}'$, the explainer first compares the number of hypotheses and deductions that might not be justified by the explainee. As $\mathcal{M}, w \vDash \Box_1 \triangle_2 F$,

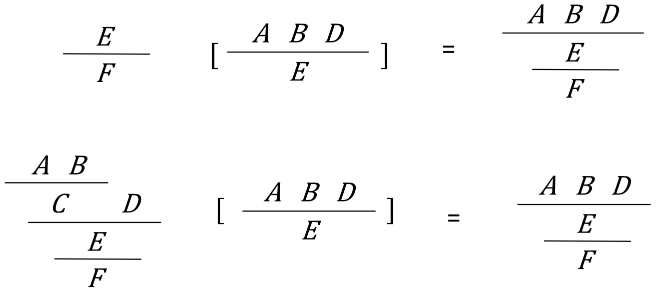

Fig. 1: Operations between explanations.

$\mathcal{E} = \mathcal{E}'[sub(\mathcal{E},F)]$, $\mathcal{E}'$ is part of $\mathcal{E}$. The explainer knows for sure that the explainee can justify F ($\mathcal{M}, w \vDash \Box_1 \triangle_2 F$), but the sub-explanation $sub(\mathcal{E},F)$ might contain hypotheses and deductions that might not be justified by the explainee. Thus, $|N^{\mathcal{M},w}(\mathcal{E})| \geq |N^{\mathcal{M},w}(\mathcal{E}')|$. If $|N^{\mathcal{M},w}(\mathcal{E})| > |N^{\mathcal{M},w}(\mathcal{E}')|$, we have $\mathcal{E} \prec^{\mathcal{M},w} \mathcal{E}'$. If $|N^{\mathcal{M},w}(\mathcal{E})| = |N^{\mathcal{M},w}(\mathcal{E}')|$, the explainee then compares the number of deduction steps in the explanations. Since $\mathcal{E}'$ is part of $\mathcal{E}$, $D(\mathcal{E}) > D(\mathcal{E}')$. So we have $\mathcal{E} \prec^{\mathcal{M},w} \mathcal{E}'$ as well.

Moreover, given an explanation $\mathcal{E}$, if the deduction in $\mathcal{E}$ can skip a step by replacing the sub-explanation for F with $\mathcal{E}''$, and the explainer is sure that the explainee can justify the skipped deduction in $\mathcal{E}''$, then the explainer will replace the sub-explanation for F with $\mathcal{E}''$, resulting a shorter explanation $\mathcal{E}'$. Skipping a step can be done by replacing one of the premises of $\mathcal{E}_a$, F, with the premises of F for constructing $\mathcal{E}''$. All the derived formulas in $\mathcal{E}''$ except the claim of $\mathcal{E}''$ are derived in the same way as in the sub-explanation for F.

Proposition 4.2. *Given a two-agent modular model $\mathcal{M}$ and a world w, for any two explanations $\mathcal{E}$ and $\mathcal{E}'$, $\mathcal{E} \prec^{\mathcal{M},w} \mathcal{E}'$ if there exists a formula $F \in D(\mathcal{E})$ and an explanation $\mathcal{E}''$ such that*

- $\mathcal{E}' = \mathcal{E}[\mathcal{E}'']$,
- *there exists $G \in Pr(\mathcal{E},F)$ s.t.*

$$Pr(\mathcal{E}'', claim(\mathcal{E}'')) = Pr(\mathcal{E},G) \cup Pr(\mathcal{E},F) \backslash \{G\};$$

and

$$\mathcal{M}, w \vDash \Box_1 \triangle_2 (\underset{Pr(\mathcal{E}'',claim(\mathcal{E}''))}{\Longrightarrow} claim(\mathcal{E}'')).$$

- *for any formula $H \in D(\mathcal{E}'') \backslash \{claim(\mathcal{E}'')\}$:*

$$Pr(\mathcal{E}'', H) = Pr(\mathcal{E}, H).$$

Proof. Given two explanations $\mathcal{E}$ and $\mathcal{E}'$, the explainer first compares the number of hypotheses and deductions that might not be justified by the explainee. Because $\mathcal{E}' = \mathcal{E}[\mathcal{E}'']$, the only difference between $\mathcal{E}$ and $\mathcal{E}'$ is the deduction for F. Since there exists

$G \in Pr(\mathcal{E}, F)$ s.t. $Pr(\mathcal{E}'', claim(\mathcal{E}'')) = Pr(\mathcal{E}, G) \cup Pr(\mathcal{E}, F) \backslash \{G\}$, which is to replace one of F's premises in $\mathcal{E}$, G, with G's premises in $\mathcal{E}$ for constructing $claim(\mathcal{E}'')$'s premises in $\mathcal{E}''$, F is derived indirectly from $Pr(\mathcal{E}'', claim(\mathcal{E}''))$ in $\mathcal{E}$, while it is done directly in $\mathcal{E}''$. As the explainer knows for sure that the explainee can justify the deduction step for $claim(\mathcal{E}'')$ in $\mathcal{E}''$ ($\mathcal{M}, w \vDash \Box_1 \triangle_2 (\underset{Pr(\mathcal{E}'', claim(\mathcal{E}''))}{\Longrightarrow} claim(\mathcal{E}''))$), but the deduction from $Pr(\mathcal{E}'', claim(\mathcal{E}''))$ to F might contain deductions that might not be justified by the explainee. Thus, $|N^{\mathcal{M},w}(\mathcal{E})| \geq |N^{\mathcal{M},w}(\mathcal{E}')|$. If $|N^{\mathcal{M},w}(\mathcal{E})| > |N^{\mathcal{M},w}(\mathcal{E}')|$, we have $\mathcal{E} <^{\mathcal{M},w} \mathcal{E}'$. If $|N^{\mathcal{M},w}(\mathcal{E})| = |N^{\mathcal{M},w}(\mathcal{E}')|$, the explainee then compares the number of deduction steps in the explanations. Since $\mathcal{E}_b$ contains fewer deduction steps than $\mathcal{E}_a$, $D(\mathcal{E}) > D(\mathcal{E}')$. So we have $\mathcal{E} <^{\mathcal{M},w} \mathcal{E}'$ as well.

Notice that Proposition 4.2 characterizes the conditions of skipping one step for making explanations, and skipping n steps can be reduced to skipping one step for n times. Both Proposition 4.1 and Proposition 4.2 highlight an intriguing point: it's not always beneficial for the explainer to provide more explanation, as the additional information might not be understood by the explainee. If the explainer knows for sure that the explainee can understand some parts of the explanation, there is no need for the explainer to elaborate further on those parts, as it could potentially cause more confusion for the explainee. Having a preference over available explanations for a claim allows an agent to select the one she prefers most.

Example 4. Fig. 1 shows explanation E/F can be combined with explanation ABD/E, resulting in explanation $ABD/E/F$. Suppose the explainer knows that the explainee can justify E as well as the deduction from E to F, namely $\mathcal{M}, w \vDash \Box_1 \triangle_2 (E \wedge (E \to F))$, but the explainer is not sure whether the explainee can justify the deduction from ABC to E. In this case, the number of hypotheses and deductions that the explainer is not sure whether the explainee can justify in E/F is less than that of $ABD/E/F$, so we have $E/F >^{\mathcal{M},w} ABD/E/F$. Suppose the explainer knows that the explainee can justify all the hypotheses and deductions in both explanations. The explainer then needs to compare the numbers of deduction steps. As $D(E/F) < D(ABD/E/F)$, we have $E/F >^{\mathcal{M},w} ABD/E/F$ as well.

5 Conversational Explanations

The explainer has incomplete information about the explainee in terms of her justified knowledge, but the explainer can gain more and more information through having feedback from the explainee on the explanations that have been announced. We first define the explainee's feedback. After the explainee hears an explanation, she can evaluate whether she can understand the explanation. So her feedback is defined inductively as a tree that is isomorphic to a given explanation so that each node in the feedback tree corresponds to a specific formula in the explanation.

Definition 5.1 (Explainees' Feedback). *Given an explanation $\mathcal{E}$, the explainee's feedback on explanation $\mathcal{E}$, denoted as $\mathcal{F}_2(\mathcal{E})$, is inductively defined as follows:*

$$\frac{\mathcal{F}_2(\Lambda_1), \ldots, \mathcal{F}_2(\Lambda_n)}{f_B}$$

where $\mathcal{F}_2(\Lambda_i) = 1$ if Λ_i is a formula of $\mathcal{L}_{\text{Prop}}$ and the explainee can justify Λ_i; $\mathcal{F}_2(\Lambda_i) = 0$ if Λ_i is a formula of $\mathcal{L}_{\text{Prop}}$ and the explainee cannot justify Λ_i; $f_B = 1$ if agent 2 understands $\mathcal{E}$; otherwise, $f_B = 0$. Given a formula F in $\mathcal{E}$, we use $\mathcal{F}_2(\mathcal{E}, F)$ to extract the value in $\mathcal{F}_2(\mathcal{E})$ that corresponds to the explainee's feedback on F. We write $\mathcal{F}(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F)$ for short if the agent is clear from the context.

As we mentioned in the previous section, if the explainee cannot understand a deduction step, then she cannot understand all the follow-up deduction steps. Thus, if there exists one node in $\mathcal{F}(\mathcal{E})$ that has value 0, all its follow-up nodes towards the root will also have value 0. Given the way we define the explainee's feedback, the explainer can learn what the explainee can justify in the explanation or understand a sub-explanation from the explainee's feedback. We first define the update of an agent's epistemic state by a truth set (this is the usual definition for announcements [21]).

Definition 5.2 (Update by a Set of Worlds). *Given a two-agent modular model $\mathcal{M}$, a subset of worlds X, and an agent i, we define $\mathcal{M}$ updated with (i, X) by*

$$\mathcal{M}|(i, X) = (Agt, W', \tilde{R}', \tilde{\ast}, \pi')$$

as follows:

- $W' = X$;
- $\tilde{R}' = \tilde{R} \cap (X \times X)$;
- $\pi' = \pi \cap X$.

We then define what the explainer can learn and interpret from the explainee's feedback on a given explanation. When the explainee returns 1 for a hypothesis F, it simply means that the explainee can justify F; when the explainee returns 0 for a hypothesis F, it simply means that the explainee cannot justify F. By Definition 3.5 on understanding an explanation, when the explainee returns 1 for a derived formula F, it means that the explainee can understand the sub-explanation for F, which also means that the explainee can justify the hypotheses as well as the deduction steps in the sub-explanation; when the explainee returns 0 for a derived formula F, it means that the explainee cannot understand the sub-explanation for F, which also means that the explainee cannot justify all the hypotheses as well as the deduction steps in the sub-explanation. Moreover, by Proposition 3.3, once the explainee understands a sub-explanation, she can then justify all the derived formulas in the sub-explanation after it is announced. Note that we assume that the explainee's feedback is truthful (otherwise it could happen that the updated model does not contain the actual world anymore, in which case we would need a truth definition similar to public announcement logic).

Definition 5.3 (Learning from Feedback). *Given a two-agent modular model $\mathcal{M}$, the update of $\mathcal{M}$ with the explainer upon receiving the explainee's feedback on the explanation $\mathcal{E}$, formally $\mathcal{M}|(1, \mathcal{F}(\mathcal{E}))$, is defined by a series of updates: for each formula F in $\mathcal{E}$ we update $\mathcal{M}$ with $(1, U_F)$ where U_F is given by*

1. *if $F \in H(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 1$, then*

$$U_F = \{w \in W \mid \mathcal{M}, w \models \triangle_2 F\};$$

2. *if $F \in H(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 0$, then*

$$U_F = \{w \in W \mid \mathcal{M}, w \vDash \neg \vartriangle_2 F\};$$

3. *if $F \in D(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 1$, then*

$$U_F = \{w \in W \mid \mathcal{M}, w \vDash \vartriangle_2 (\bigwedge_{A \in H(sub(\mathcal{E},F))} A \wedge \bigwedge_{B \in D(sub(\mathcal{E},F))} \underset{Pr(\mathcal{E},B)}{\Longrightarrow} B \wedge B)\};$$

4. *if $F \in D(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 0$, then*

$$U_F = \{w \in W \mid \mathcal{M}, w \vDash \neg \vartriangle_2 (\bigwedge_{A \in H(sub(\mathcal{E},F))} A \wedge \bigwedge_{B \in D(sub(\mathcal{E},F))} \underset{Pr(\mathcal{E},B)}{\Longrightarrow} B)\};$$

5. *otherwise, $U_F = W$.*

Even though the update is done in series, the order of the updates does not matter. In other words, one could just do one update with the intersection of all U_F's whose conditions apply. The above definition gives rise to the following results.

Proposition 5.1. *Given a two-agent modular model $\mathcal{M}$, the updated model $\mathcal{M}|(1, \mathcal{F}(\mathcal{E}))$ with the explainer upon receiving the explainee's feedback on the explanation $\mathcal{E}$ is still a two-agent modular model.*

Proof. The update approach defined in Definition 5.2 remove states from the system. We first check whether the R relation in the updated model is still reflexive and transitive. Suppose it is not reflexive any more, which means that there exists a world $w \in W'$ such that $(w, w) \notin R'_i$. However, since $w \in W$, it is also the case that $(w, w) \notin R_i$, making R_i not reflexive in the original model. Contradiction! Suppose it is not transitive, which means that there exists worlds w_1, w_2 and $w_3 \in W'$ such that $(w_1, w_2) \in R'_i$ and $(w_2, w_3) \in R'_i$, but $(w_1, w_3) \notin R'_i$. However, since worlds w_1, w_2 and $w_3 \in W$, it is also the case that $(w_1, w_2) \in R_i$ and $(w_2, w_3) \in R_i$, but $(w_1, w_3) \notin R_i$, making R_i not transitive in the original model. Contradiction! Next, we check whether the updated model still has validity JYK. As the update approach removes states from the model, agents' knowledge get expanded after update. So if agent i knows F in the original model, it is still the case in the updated model. Suppose agent i uses t to justify F in world w in the original model. By validity JYK, agent i also knows F in world w in the original model. As announcements are always truthful, the updated model still contains the actual world w. So in the updated model, it is still the case that agent i uses t to justify F and knows F in world w.

We then extend our language $\mathcal{L}_\mathrm{J}$ with new formulas of the form $[j : \mathcal{F}(\mathcal{E})]F$, read as "$F$ is true after agent j hears feedback $\mathcal{F}(\mathcal{E})$". We set

$$\mathcal{M}, w \vDash [j : \mathcal{F}(\mathcal{E})]F \quad \text{iff} \quad \mathcal{M}|(j, \mathcal{F}(\mathcal{E})), w \vDash F.$$

This formula allows to express agent j's updated epistemic state after hearing feedback on an explanation. The result from the update is quite straightforward. However, one should notice that, the explainer knows that the explainee cannot justify a deduction step only when the explainee returns 1 for the premises but 0 for the derived formula. This is because when agent 2 returns 0 for a derived formula as well as some of its premises, the explainer cannot understand the sub-explanation for the derived formula due to its premises *or* the deduction step in between, making the explainer not able to tell whether the explainee can justify the deduction or not. In such a case, the explainer will ignore the feedback with respect to this and all the follow-up deduction steps.

Proposition 5.2. *Given a two-agent modular model $\mathcal{M}$ and a world w, the explainer hears feedback $\mathcal{F}(\mathcal{E})$ from the explainee on explanation $\mathcal{E}$ in world w, for any formula F in $\mathcal{E}$,*

- *if $F \in H(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 1$, then*

$$\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1\, \triangle_2 F$$

- *if $F \in H(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 0$, then*

$$\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1 \neg \triangle_2\, F$$

- *if $F \in D(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 1$, then*

$$\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1\, \triangle_2(\underset{Pr(\mathcal{E},F)}{\Longrightarrow} F).$$

- *if $F \in D(\mathcal{E})$, $\mathcal{F}(\mathcal{E}, F) = 0$ and for any $P \in Pr(\mathcal{E}, F)$ it is the case that $\mathcal{F}(\mathcal{E}, F) = 1$, then*

$$\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1 \neg \triangle_2\, (\underset{Pr(\mathcal{E},F)}{\Longrightarrow} F).$$

Proof. If $F \in H(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 1$, then only worlds that satisfy $\triangle_2 F$ will be kept, which results in $\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}]\Box_1 \triangle_2 F$. If $F \in H(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 0$, then only worlds that satisfy $\neg \triangle_2 F$ will be kept, which results in $\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1 \neg \triangle_2\, F$. If $F \in D(\mathcal{E})$ and $\mathcal{F}(\mathcal{E}, F) = 1$, worlds where the explainee can justify all the hypotheses and deductions in sub-explanation $sub(\mathcal{E}, F)$ are kept, resulting in $\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1\, \triangle_2(\underset{Pr(\mathcal{E},F)}{\Longrightarrow} F)$. As for the final one, if $F \in D(\mathcal{E})$, $\mathcal{F}(\mathcal{E}, F) = 0$ and for any $P \in Pr(\mathcal{E}, F)$ it is the case that $\mathcal{F}(\mathcal{E}, F) = 1$, then worlds that satisfy $\neg \triangle_2\, (\bigwedge_{A \in H(sub(\mathcal{E},F))} A \wedge \bigwedge_{B \in D(sub(\mathcal{E},F))} \underset{Pr(\mathcal{E},B)}{\Longrightarrow} B)$ are kept. Notice that

$$\neg \triangle_2\, (\bigwedge_{A \in H(sub(\mathcal{E},F))} A \wedge \bigwedge_{B \in D(sub(\mathcal{E},F))} \underset{Pr(\mathcal{E},B)}{\Longrightarrow} B)$$

is equivalent to

$$\bigvee_{A \in H(sub(\mathcal{E},F))} \neg \triangle_2\, A \vee \bigvee_{B \in D(sub(\mathcal{E},F))} \neg \triangle_2 \underset{Pr(\mathcal{E},B)}{\Longrightarrow} B).$$

Since for any $P \in Pr(\mathcal{E}, F)$ we have $\mathcal{F}(\mathcal{E}, F) = 1$, the disjuncts $\bigvee_{A \in H(sub(\mathcal{E},F))} \neg \triangle_2 A$ as well as $\bigvee_{B \in D(sub(\mathcal{E},F)) \setminus F} \neg \triangle_2 \underset{Pr(\mathcal{E},B)}{\Longrightarrow} B)$ can be eliminated. Thus, it is the case that $\mathcal{M}, w \vDash [1 : \mathcal{F}(\mathcal{E})][2 : \mathcal{E}] \,\Box_1 \neg \triangle_2\, (\underset{Pr(\mathcal{E},F)}{\Longrightarrow} F)$.

Using the background information about the explainee, the explainer can further explain what the explainee cannot justify. However, agent 1 should always remember that its goal is to answer the initial question from the explainee instead of infinitely explaining what the explainee cannot justify in the previous explanation. All of these require the explainer to memorize the conversation with the explainee. We first define the notion of conversation histories between two agents, which always starts with a question and is followed by an explanation from the explainer and feedback from the explainee in an alternating way.

Definition 5.4 (Conversation Histories). *A finite conversation history between the explainer and the explainee for explaining a propositional formula F is of the form*

$$\eta = (1 :?F)(2 : \mathcal{E}_1)(1 : \mathcal{F}(\mathcal{E}_1))\cdots(2 : \mathcal{E}_n)(1 : \mathcal{F}(\mathcal{E}_n)),$$

where $\mathcal{E}_1\cdots\mathcal{E}_n$ are explanations made by the explainer to the explainee. We use

$$pre(\eta, k) = (1 :?F)(2 : \mathcal{E}_1)(1 : \mathcal{F}(\mathcal{E}_1))\cdots(2 : \mathcal{E}_k)(1 : \mathcal{F}(\mathcal{E}_k))$$

to denote the prefix of η up to the k^{th} position.

Given a conversation history, the explainer can decide the explanation to be announced. The decision-making is formalized as a function $\mathcal{E}^*(\eta)$ that takes a conservation history as an input. The explainer needs to evaluate whether it is more worthy to further explain what the explainee cannot justify in the previous explanation, or find another way to explain the initial question, given the current information about the explainee. Let us first use $why(\mathcal{F}(\mathcal{E}))$ to denote the set of formulas that are supposed to have further explanation given feedback $\mathcal{F}(\mathcal{E})$.

$$\begin{aligned} why(\mathcal{F}(\mathcal{E})) = \{P \in \mathcal{E} \mid \mathcal{F}(\mathcal{E}, P) = 0 \text{ and}& \\ (P \in H(\mathcal{E}) \text{ or } \mathcal{F}(\mathcal{E}, G) = 1 \text{ for all } G \in Pr(\mathcal{E}, P)).\}& \end{aligned}$$

The set $why(\mathcal{F}(\mathcal{E}))$ contains formulas in explanation $\mathcal{E}$ on which the explainee asks for further explanations and that are either hypotheses or derived formula whose premises are justified by the explainee. The rest of the unjustified formulas explanation $\mathcal{E}$ should not be explained, because it is not clear whether the explainee cannot justify their premises or deductions. As we mentioned before, the explainer's goal is to answer the initial question from the explainee. Thus, when looking for the most preferred explanations, the explainer should not only consider the explanations for the questions that were asked right now, but also the explanations for the initial question.

Definition 5.5 (Most Preferred Explanations). *Assume that we are given a two-agent modular model $\mathcal{M}$, a world w, and a conversation history*

$$\eta = (1 :?F)(2 : \mathcal{E}_1)(1 : \mathcal{F}(\mathcal{E}_1))\cdots(2 : \mathcal{E}_n)(1 : \mathcal{F}(\mathcal{E}_n)).$$

The set of the most preferred explanations with respect to η is given by a function $\mathcal{E}^*(\eta)$, which is defined as follows.

$$\mathcal{M}' := \mathcal{M}|(2:\mathcal{E}_1)|(1:\mathcal{F}(\mathcal{E}_1))|\cdots|(2:\mathcal{E}_n)|(1:\mathcal{F}(\mathcal{E}_n))$$

$$X := \lambda^{\mathcal{M}',w}(\varnothing, F) \cup \bigcup_{G\in why(\mathcal{F}(2,\mathcal{E}_n))} \lambda^{\mathcal{M}',w}(Pr(\mathcal{E}_n,G),G)$$

$$\mathcal{E}^*(\eta) := \{\mathcal{E}\in X \mid \textit{for any } \mathcal{E}'\in X \textit{ it is the case that } \mathcal{E}' \lesssim^{\mathcal{M}',w} \mathcal{E}\}$$

In words, given a conversation history η, the explainer evaluates explanations from the set $\lambda^{\mathcal{M}',w}(F)$, which is the set of available explanations for the initial question regarding F, and $\bigcup_{G\in why(\mathcal{F}(\mathcal{E}_n))} \lambda^{\mathcal{M}',w}(Pr(\mathcal{E}_n,G),G)$, which is the set of available explanations for further explaining what the explainee cannot undFerstand in $\mathcal{E}_n$, and chooses the one that is most preferred based on the principles in Definition 4.2. For example, if agent 1 finds that it is too time-consuming to make agent 2 understand $\mathcal{E}_n$, because any further explanations contain too many deduction steps, then the explainer might prefer explaining the claim F in another simple way, if there exists one. It is also important to stress that the evaluation is made with respect to model $\mathcal{M}' = \mathcal{M}|(2:\mathcal{E}_1)|(1:\mathcal{F}(\mathcal{E}_1))|\cdots|(2:\mathcal{E}_n)|(1:\mathcal{F}(\mathcal{E}_n))$, which means that the explainer considers the latest information about the explainee that she can infer from conversation η.

Since explanations are conducted conversationally, it is of great importance for the conversation to terminate. The conversation can terminate due to two reasons: either the explainee has justified the initial claim and thus does not ask any more questions, or the explainer cannot explain more. For sure, we would like the first one to occur. The following proposition expresses that our explanation approach ensures this desired property if there exists an explanation that the explainee understands and the explainer is aware of, and the explanation can be constructed through a series of replacement operations over the announced explanations in the conversation history.

Theorem 5.1. *Given a two-agent modular model $\mathcal{M}$ and a world w, if there exists an explanation $\mathcal{E}$ such that $claim(\mathcal{E}) = F$, the explainee understands $\mathcal{E}$ and the explainer is aware of $\mathcal{E}$ in world w, then there exists a conversation history*

$$\eta = (1:?F)(2:\mathcal{E}_1)(1:\mathcal{F}(\mathcal{E}_1))\cdots(2:\mathcal{E}_n)(1:\mathcal{F}(\mathcal{E}_n)),$$

where $\mathcal{E}_i \in \mathcal{E}^*(pre(\eta, i-1))$, such that

1. *$\mathcal{M}, w \models [2:\mathcal{E}_n]\cdots[1:\mathcal{F}(\mathcal{E}_1)][2:\mathcal{E}_1][\![t]\!]_2 F$, where $t \in \mathrm{Gt}$ is a derived term that the explainee constructs with respect to $\mathcal{E}$;*
2. *if $n = 1$, $\mathcal{E} = \mathcal{E}_1$; if $n > 1$, $\mathcal{E} = \mathcal{E}_1[\mathcal{E}_2]\cdots[\mathcal{E}_n]$;*
3. *all the nodes in $\mathcal{F}(\mathcal{E}_n)$ have value 1.*

Proof. 1. We first regard all the explanations that the explainer is aware of for the claim F as a tree that is rooted at F, denoted as T_F. Every time the explainee provides feedback to the explainer's previous explanation, the explainer can have more information about agent 2 (the formulas and deductions that the explainee can justify). With this information, agent 2 can remove certain formulas and deductions between formulas from T_F so that the explanations containing formulas and deductions that the explainee cannot justify are no longer available for the explainer to the explainee. Since the explainee's feedback is always truthful, the explainer will not remove formulas and deductions that the explainee can justify from T_F. Thus, if there exists an explanation $\mathcal{E}$ with $claim(\mathcal{E}) = F$, the explainee understands $\mathcal{E}$ and the explainer is aware of $\mathcal{E}$, then $\mathcal{E}$ will always stay in T_F. Using the approach defined in Definition 5.5 allows the explainer to look back at the explanations with $claim(\mathcal{E}) = F$ in each round of explanation selection. Recall that we have constraints on the evidence function: terms that can be used to justify a formula F are finite, and the formulas that a given term can justify are finite. These two constraints ensure that a formula F has finitely many explanations (if there exists). As $\mathcal{E}$ will always stay in T_F, and the explanations that can be used to justify F are finite, $\mathcal{E}$ will be found by the explainer at some point. Once the explainee constructs or hears $\mathcal{E}$ that she can understand, she can justify F with a ground term t. Therefore, if there exists an explanation $\mathcal{E}$ with $claim(\mathcal{E}) = F$, the explainee understands $\mathcal{E}$ and the explainer is aware of $\mathcal{E}$, there exists η such that after η the explainee can justify F.

2. If $n = 1$, $\mathcal{E} = \mathcal{E}_1$ trivially holds. The following will prove $\mathcal{E} = \mathcal{E}_1[\mathcal{E}_2]\cdots[\mathcal{E}_n]$ for $n > 1$. The explanation that the explainer chooses to announce is either to further explain what the explainee cannot justify in the previous explanation, or to explain F in a completely different way. The above has proved that $\mathcal{E}$ can be found by the explainee. If $\mathcal{E} = \mathcal{E}_n$, then $\mathcal{E}_n$ completely replaces the explanation that has been built previously, i.e., there exists m with $1 \leq m < n$ such that $\mathcal{E} = \mathcal{E}_m[\mathcal{E}_{m+1}]\cdots[\mathcal{E}_{n-1}][\mathcal{E}_n]$, where $\mathcal{E}_m$ has claim F and is also constructed in the previous steps. If $\mathcal{E} \neq \mathcal{E}_n$, then $\mathcal{E}$ is built together with previous steps, i.e., there exists m with $1 \leq m \leq n$ such that $\mathcal{E} = \mathcal{E}_m[\mathcal{E}_{m+1}]\cdots[\mathcal{E}_n]$, where $\mathcal{E}_m$ has claim F and is also constructed in the previous steps. As η is finite, the above backward reasoning will eventually reach $\mathcal{E}_1$. Hence, given the conversation history η, $\mathcal{E} = \mathcal{E}_1[\mathcal{E}_2]\cdots[\mathcal{E}_n]$.
3. As $\mathcal{E}$ is constructed through $[\mathcal{E}_i]$, $\mathcal{E}_n$ is part of $\mathcal{E}$ or equal to $\mathcal{E}$. If some of the nodes in $\mathcal{F}(\mathcal{E}_n)$ have value 0, which means that the explainee has questions about $\mathcal{E}_n$, then the explainee cannot construct a ground derived term for F with respect to $\mathcal{E}$ due to her learning approach in Definition 3.4, which contradicts the previous conclusion. So all of the nodes in $\mathcal{F}(\mathcal{E}_n)$ have value 1.

Example 5. We illustrate our approach using the example that was mentioned in the introduction. A user u asks a chatbot c why she should drink more water. Assume that the chatbot has no information about the user except the user being sick. She thereby announces explanation $\mathcal{E}$ to the user, which is "being sick can lead to fluid loss, so drinking more water helps replenish these losses", formalized as $\mathcal{E} = sick/fluid\_loss/drink\_water$. However, the user replies to the chatbot with $\mathcal{F}(\mathcal{E}) = 1/0/0$, making the chatbot knows that the user can justify that she is sick but cannot justify the deduction from $sick$ to $fluid\_loss$,

$$\mathcal{M}, w \vDash [c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}] \, \Box_c \, \triangle_u sick,$$
$$\mathcal{M}, w \vDash [c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}] \, \Box_c \neg \triangle_u (sick \to fluid\_loss).$$

These express the chatbot's knowledge about the user's background after the chatbot hears her user's feedback $\mathcal{F}(\mathcal{E})$. Next, the chatbot needs to decide what to explain.

Regarding $\mathcal{E}$, the chatbot can further explain why being sick can lead to fluid loss, namely $why(\mathcal{F}(\mathcal{E})) = \{fluid\_loss\}$. Besides, the chatbot can also explain why to drink more water when being sick in another way. Suppose the chatbot recognizes that given her knowledge about the user's background it is too complicated to explain why being sick can lead to fluid loss, namely $\mathcal{E}' = sick/\cdots/fluid\_loss$, but she can simply tell the user that being sick can make you thirsty, so you should drink more water, formalized as $\mathcal{E}'' = sick/thirsty/drink\_water$. Given the conversation history

$$\eta = (c : ?drink\_water)(u : \mathcal{E})(c : \mathcal{F}(\mathcal{E})),$$

the chatbot knows that the user can justify the deduction from $sick$ to $thirsty$ and from $thirsty$ to $drink\_water$ in $\mathcal{E}''$. Using the criteria in Definition 4.2, the chatbot specifies the preference over $\mathcal{E}'$ and $\mathcal{E}''$, and gets $\mathcal{E}' <^{\mathcal{M}',w} \mathcal{E}''$, where $\mathcal{M}' = \mathcal{M}|(u : \mathcal{E})|(c : \mathcal{F}(\mathcal{E}))$. Thus, the chatbot announces $\mathcal{E}''$ to the user. Since the chatbot's knowledge about the user is always true, the user can justify all parts in $\mathcal{E}''$ and thus understand $\mathcal{E}''$. More precisely, assume that $t, s, r \in \mathrm{Gt}$, and

$$\begin{aligned}
&\mathcal{M}, w \vDash [c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}][\![t]\!]_u sick,\\
&\mathcal{M}, w \vDash [c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}][\![s]\!]_u (sick \to thirsty),\\
&\mathcal{M}, w \vDash [c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}][\![r]\!]_u (thirsty \to drink\_water).
\end{aligned}$$

The user then constructs term $s \cdot t$ for $thirsty$ and term $r \cdot (s \cdot t)$ for $drink\_water$ with respect to $\mathcal{E}''$, and gains more justified knowledge accordingly after hearing $\mathcal{E}''$,

$$\begin{aligned}
&\mathcal{M}, w \vDash [u : \mathcal{E}''][c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}][\![s \cdot t]\!]_u thirsty,\\
&\mathcal{M}, w \vDash [u : \mathcal{E}''][c : \mathcal{F}(\mathcal{E})][u : \mathcal{E}][\![r \cdot (s \cdot t)]\!]_u drink\_water.
\end{aligned}$$

Therefore, the user's feedback on $\mathcal{E}''$ is $\mathcal{F}(\mathcal{E}'') = 1/1/1$. After the chatbot hears this feedback, the conversation terminates.

6 Related Work

Within philosophy and psychology, there has been a lot of debate about what an explanation is, but accounts of explanation are commonly associated with causality [9,23,31]. Miller argues in her seminal paper that explanation can be defined as two processes and one product [19]. Explanation is a process of abductive inference for 'filling the gaps' between a given event and the causes for the event. In this paper, agents might not be able to derive all the logical consequences of their justified knowledge, and explanation helps to derive the missing parts. The explanation that results from the above abductive inference process is the product of the cognitive explanation process. Explanation is also a process of transferring knowledge between explainer and explainee, in which the goal is that the explainee has enough information to understand the causes of the event. This social aspect of explanation allows us to use the conversational approach to provide personalized explanations. In philosophy, Carl Hempel articulated the first theory of explanation. Hempel claimed that there are two types of explanation, what she called 'deductive-nomological' (DN) and 'inductive-statistical' (IS) respectively [10].

For Hempel, DN explanations were always to be preferred to IS explanations, because she understood the concept of explanation as a purely logical concept, i.e., there is a logical relationship between premises and conclusion in explanation, which is consistent with the way we define explanation.

Although there is a solid psychological and philosophical background of explanation, formal accounts of explanation have just been proposed largely in the guise of so-called Explainable AI. Shvo et al. achieve that in terms of the knowledge of agents and the mechanism by which agents revise their knowledge given possible explanations [25]. Vasileiou et al. define the notion of logic-based explanations in the context of model reconciliation problems [28].

In the area of logic, we find the logic of knowing why that has been introduced by Xu et al. [32]. They extend the language of epistemic logic with new modalities to express *agent i knows why φ*. These modalities are part of Wang's project [29] of developing a new generation of epistemic logics that go beyond knowing that. The logic of knowing why has been extended by public announcement operators for announcing formulas and also for announcing reasons [20]. Based on the logic of knowing why, Wei [30] presents a logic of understanding why that is based on two types of explanations: whereas knowing why requires knowing horizontal explanations, understanding why additionally requires vertical explanations. In philosophy, grounding is considered to be a form of metaphysical explanation. For example, the existence of sets members is thought to explain the existence of sets. The most influential logic of ground is the weak ground Kit Fine develops [7], and Adam Lovett then did follow-up research by proposing a logic of strict ground [17]. An explanation can also be expressed in the form "A because B", and Benjamin Schnieder develops a logic for 'because' based on systematic connections between 'because' and the truth-functional connectives [24].

7 Conclusion

It is important for our AI systems to provide personalized explanations to users that are relevant to them and align with their background knowledge. A conversation between explainers and explainees not only allows explainers to obtain the explainees' background but also allows explainers to tailor their explanations so that explainees can better understand the explanations. In this paper, we have proposed an approach that allows an explainer to tailor and communicate personalized explanations to an explainee through having consecutive conversations with the explainee. It is built on the idea that the explainee understands an explanation if and only if she can justify all formulas in the explanation. In a conversation for explanations, the explainee provides her feedback on the explanation that has just been announced, while the explainer interprets the explainee's background from the feedback and then selects an explanation for announcement given what has learned about the explainee. We have proved that the conversation will terminate due to the explainee's justification of the initial claim as long as there exists an explanation for the initial claim that the explainee understands and the explainer is aware of. In the future, we would like to extend our approach with another dimension for evaluating explanations: the *acceptance* of explanations. The explanation that is selected by the explainer should be not only understood, but also accepted, by the explainee. For example, a policeman does not accept an explanation for over-speed driving due to heading for a party. For this part of study, our framework needs to be enriched with evaluation standards such as values and personal norms. Moreover, we can also study personalized explanations from the nonmonotonic perspective. Our modal can deal with cases where there is disagreement among the agents about deductions: an agent perceives exceptions from a deduction made by another agent, hence requiring argumentation between the agents. Another direction for future research can be to provide axiomatization for our model, which should characterize the two updates of both agents learning from explanations and learning from feedback apart from justification logic and modal logic in S4.

References

1. Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.
2. Sergei Artemov. Explicit provability and constructive semantics. *Bulletin of Symbolic Logic*, 7(1):1–36, March 2001.
3. Sergei Artemov. The ontology of justifications in the logical setting. *Studia Logica*, 100(1–2):17–30, 2012.
4. Sergei Artemov and Melvin Fitting. *Justification Logic: Reasoning with Reasons*. Cambridge University Press, 2019.
5. Sergei Artemov, Melvin Fitting, and Thomas Studer. Justification Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2024 edition, 2011.
6. Sergei Artemov and Elena Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15(6):1059–1073, 2005.
7. Kit Fine. Guide to ground. *Metaphysical grounding: Understanding the structure of reality*, 37:40, 2012.
8. Melvin Fitting. The logic of proofs, semantically. *APAL*, 132(1):1–25, February 2005.
9. Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 2005.
10. Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.
11. Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990.
12. Denis J Hilton. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4):273–308, 1996.
13. Joseph MF Jaspars and Denis J Hilton. Mental models of causal reasoning. 1988.
14. Roman Kuznets and Thomas Studer. Justifications, ontology, and conservativity. In Thomas Bolander, Torben Braüner, Silvio Ghilardi, and Lawrence Moss, editors, *Advances in Modal Logic, Volume 9*, pages 437–458. College Publications, 2012.
15. Roman Kuznets and Thomas Studer. *Logics of Proofs and Justifications*. College Publications, 2019.
16. Eveline Lehmann and Thomas Studer. Subset models for justification logic. In R. Iemhoff, M. Moortgat, and R. de Queiroz, editors, *Logic, Language, Information, and Computation - WoLLIC 2019*, pages 433–449. Springer, 2019.

17. Adam Lovett. The logic of ground. *Journal of Philosophical Logic*, 49(1):13–49, 2020.
18. Jieting Luo, Thomas Studer, and Mehdi Dastani. Providing personalized explanations: A conversational approach. In *International Conference on Logic and Argumentation*, pages 121–137. Springer, 2023.
19. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
20. Nicholas Pischke. Dynamic extensions for the logic of knowing why with public announcements of formulas, 2018.
21. Jan Plaza. Logics of public communications. *Synthese*, 158(2):165–179, 2007.
22. Stephen J Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3):429, 1993.
23. Wesley C Salmon. *Four decades of scientific explanation*. University of Pittsburgh press, 2006.
24. Benjamin Schnieder. A logic for 'because'. *The Review of Symbolic Logic*, 4(3):445–465, 2011.
25. Maayan Shvo, Toryn Q Klassen, and Sheila A McIlraith. 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 2*, pages 75–93. Springer, 2020.
26. Paul Thagard. Explanatory coherence. *Behavioral and brain sciences*, 12(3):435–467, 1989.
27. Chun-Hua Tsai and Peter Brusilovsky. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction*, 31(3):591–627, 2021.
28. Stylianos Loukas Vasileiou, William Yeoh, Tran Cao Son, Ashwin Kumar, Michael Cashmore, and Dianele Magazzeni. A logic-based explanation generation framework for classical and hybrid planning problems. *Journal of Artificial Intelligence Research*, 73:1473–1534, 2022.
29. Yanjing Wang. Beyond knowing that: a new generation of epistemic logics. *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 499–533, 2018.
30. Yu Wei. A logical framework for understanding why. In Alexandra Pavlova, Mina Young Pedersen, and Raffaella Bernardi, editors, *Selected Reflections in Language, Logic, and Information*, pages 203–220. Springer, 2024.
31. James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.
32. Chao Xu, Yanjing Wang, and Thomas Studer. A logic of knowing why. *Synthese*, 198:1259–1285, 2021.