A Dialogue System for identifying Disagreements in Deductive Reasoning

Yifan is working on a dialogue based system for explaining reasoning in rules-based systems. We have worked through various versions of this with a Prolog implementation, starting out treating proofs as acyclic graphs with (we hoped) some kind of completeness property (i.e., the graph would capture all necessary information about the proof) and moving slowing to an idea around proof trees with Prolog-style negation as failure.

Our starting point is two "players" (assumed to be some deductive system and a user). Each possesses some set of facts (F) and a set of rules (R) of the form *if ... then ...* and use these to deduce whether some conclusion, c, is true or false. Deductions are represented as trees. Each node in the tree is labelled with the fact or conclusion deduced (represented as a first-order formula) and either the label `initial` indicating an initial fact, `unprovable` meaning the fact or conclusion can not be proved (and so is deduced to be false), or the rule that was used to deduce term labelling the node.

When the players disagree about some conclusion they engage in a dialogue. Each player can ask *why* a particular node is believed in which case they are informed that it was either an initial fact, or it was deduced from its parent nodes using the rule. A player can also ask *why not* questions – i.e., why does some conclusion is unprovable. In this case the system turns this around and asks them why they believe that it is.

The intention is that through this dialogue the user can understand either that the system did (or did not) believe some initial fact that the user does not (or does) believe or that the system and user's rules differ.

Note we assume that both players reason correctly and do not consider the case where mistakes are made in the reasoning process.

This note aims at trying to establish a formal definition of the dialogue process and formal properties of that definition (particularly that it terminates with a disagreement found). Our starting point was [Dennis and Oren, 2021] which does something similar in the context of BDI-style reasoning.

1 Preliminaries

We have some language of terms, $\mathcal{L}$, defined in a standard way.

- We have a set of labels L which include two special labels, *initial* and *unprovable*.
- We have sets of initial facts, F, these consist of positive literals in $\mathcal{L}$.
- We have sets of rules, R. A rule is a horn clause consisting of a non-empty set of literals in $\mathcal{L}$ (the antecedants, A), and a consequent, a positive literal $C \in \mathcal{L}$, and a label $l \in L \backslash \{initial, unprovable\}$. We write a rule as $l : A \rightarrow C$. We assume that labels in R are unique – i.e., there is only one rule labelled l in any set of rules, R. Note that we don't need to label rules for our system to work, but labels are a useful convenience when referring to rules.

Definition 1. (Positive and Negative Literals in a set) *For a set of literals, A,*

$$pos(A) = \{l : l \text{ is a positive literal } \wedge l \in A\}$$

(the set of positive literals appearing in A).

For a set of literals, A,

$$neg(A) = \{l : l \text{ is a positive literal } \wedge \neg l \in A\}$$

(the set of literals appearing negatively in A).

1.1 Proof Trees

Definition 2. (Proof Tree) *A proof tree is a directed rooted tree written $\langle N, E\rangle$, where N is the set of nodes in the tree, and $E \subseteq N \times N$ is the set of edges. Each node $n \in N$ consists of a pair of a ground positive literal, $t \in \mathcal{L}$ and a label, $l \in L$ written (t, l). An edge between two nodes n_1 and n_2 is written as $n_1 \mapsto n_2$.*

We use standard terminology so the root of a proof tree is the single node, n such that there is no edge $n' \mapsto n$. A node, n's, parent nodes are the set of nodes, n', such that there exists an edge $n \mapsto n'$. A node, n's, parent trees are the set of sub-trees with a parent of n as their root.

If (t, l) is the root node of a tree, then we refer to t as the root term *of the tree.*

*Notes in this series are for ϵ-baked ideas, for $1 \geq \epsilon \geq 0$. Only exceptionally should they be cited.

Definition 3. Provable and Unprovable in T *If* $\langle N, E\rangle = T$ *is a proof tree and* t *is a ground positive literal in* $\mathcal{L}$*. We say:*

- t *is* provable *in* T *iff there exists a node* $(t, l) \in N$ *such that* $l \neq unprovable$*.*
- t *is* unprovable *in* T *iff* $(t, unprovable) \in N$*.*
- t *is* undecided *in* T *iff there is no node* $(t, l) \in N$*.*

Definition 4. (Proof Tree for F **and** R**)** *A* proof tree*,* T*, for a set of facts,* F*, and rules,* R *is defined recursively as follows:*

- $\langle\{(t, initial)\}, \emptyset\rangle$ *is proof tree for* F *and* R *iff* $t \in F$
- $\langle\{(t, unprovable)\}, \emptyset\rangle$*, is a proof tree for* F *and* R *iff no proof tree,* T'*, for* F *and* R *exists such that* t *is provable in* T'
- *If* $E \neq \emptyset$ *then a proof tree* $T = \langle N, E\rangle$ *with root node* (t, l) *is a proof tree for* F *and* R *iff:*
  - *The parent trees of* (t, l) *are all proof trees for* F *and* R
  - *There exists a rule,* $l : A \rightarrow C \in R$ *and a substitution,* θ *for the free variables in* A *and* C *such that* $C\theta = t$ *and* $t \notin F$*.*
  - *If* (t', l') *is a parent of* (t, l) *in* T *then either*
    - $\exists t_i \in pos(A).\ t_i\theta = t'$ *and* $l' \neq unprovable$ *or,*
    - $\exists t_i \in neg(A).\ t_i\theta = t'$ *and* $l' = unprovable$*.*
  - *For all* t' *such that* $t' \in A\theta$ *there exists a unique label,* l' *such that* (t', l') *is a parent node of* (t, l) *in* T*. Note that because* $A \neq \emptyset$ *this means that if any proof tree,* T*, for* F *and* R *that contains a node* (t, l) *where* $l \notin \{initial, unprovable\}$ *contains more than one node and at least one edge.*

We note (though we don't prove) that a proof tree with some statement t at its root (either as a provable or unprovable statement) can be constructed from F and R by standard backwards reasoning in a Prolog style.

I'm going to stop talking about substitutions, θ etc., from this point. They make everything harder to read and I don't think they make much difference to correctness.

We will also assume that we are only dealing with sets of facts and rules that always terminate when attempting to find a proof tree with some term t as the root.

Theorem 1. *If there exists a proof tree,* T*, for* F *and* R *and a term* t *which is unprovable in* T *then there is no proof tree for* F *and* R *in which* t *is provable.*

Proof. This is a very simple proof by induction. In the base case, if t is unprovable in T then $T = \langle\{(t, unprovable)\}, \emptyset\rangle$ and there is no proof tree for F and R in which t is provable by definition. In the step case t is unprovable in some T with root node (t, l) where $l : A \rightarrow C \in R$. Since no rule is labelled *unprovable* by definition, t must be unprovable in one of the sub-trees of T and so, by the induction hypothesis, there is no proof tree for F and R in which t is provable. □

Theorem 2. *If there exists a proof tree,* $\langle N, E\rangle$*, for* F *and* R *and a node* $(t, l) \in N$ *where* $l \notin \{initial, unprovable\}$ *then there exists a rule* $l : A \rightarrow C \in R$*.*

Proof. Again this is a very simple proof by induction. In the base case single node trees are labelled with either *initial* or *unprovable* so the theorem is trivially true. Otherwise in the step case, if the root of the tree is the (t, l) then there exists a rule, $l : A \rightarrow C \in R$ by definition otherwise (t, l) is in one of the sub-trees and the existence of a rule $l : A \rightarrow C \in R$ follows from the induction hypothesis. □

2 Problem Statement

We need to formalise the idea of a disagreement between the user and the reasoning system. We will simplify the problem to identifying a *cause* of a difference between two proof trees (this is obviously a massive simplification of why a user might want more information about a reasoning system's deduction, but it is a solid first step).

Definition 5. *We formalise a deduction as a tuple* $\mathcal{D}(F, R, \mathcal{T})$ *where* F *is a set of initial facts,* R *a set of rules and* $\mathcal{T}$ *is a set of proof trees for* F *and* R*. We will refer to* F *as the* deduction facts*,* R *as the* deduction rules *and* $\mathcal{T}$ *as the* deduction trees*.*

Our problem is, given two initial deductions $\mathcal{D}(F_1, R_1, \{T_1\}) \neq \mathcal{D}(F_2, R_2, \{T_2\})$ such that T_1 and T_2 have the same root term but that root term is provable in one and unprovable in the other, can we identify at least one fact, t such that $t \in F_1$ and $t \notin F_2$ (or vice versa) or at least one rule r such that $r \in R_1$ and $r \notin R_2$. Note that we can trivially identify the differences if we have full access to F_1, F_2, R_1, R_2 etc., So our problem is whether we can find out by an incremental dialogue process where the starting information is simply a root term that is provable in one deduction tree and not in the other. We assume that rules with the same label in R_1 and R_2 are identical – i.e., if $l : A_1 \rightarrow C_1 \in R_1$ and $l : A_2 \rightarrow C_2 \in R_2$ then $A_1 = A_2$ and $C_1 = C_2$. This means we can use rule labels as proxies for the rules themselves rather than having to match antecedants and consequents.

3 A Dialogue

A dialogue is a sequence of moves taken by two players. P_1 knows all the information in $D_0^1 = \mathcal{D}(F_1, R_1, \{T_1\})$ while P_2 knows all the information in $D_0^2 = \mathcal{D}(F_2, R_2, \{T_2\})$. The two players gradually build up a model of how the other player has reasoned. This model consists of four sets Y_{ij}, F_{ij}, N_{ij} and YR_{ij}:

- Y_{ij} consists of terms t that P_i has established that P_j believes to be true. We refer to Y_{ij} as the *opponent belief set.*
- F_{ij} consists of terms t that P_i has established that P_j had as an initial fact. Note that $F_{ij} \subseteq Y_{ij}$. We refer to F_{ij} as the *opponent fact set.*
- N_{ij} consists of terms t that P_i has established that P_j does not believe to be true. We refer to N_{ij} as the *opponent disbelief set.*
- YR_{ij} consists of rule labels l that P_i has established label on of P_j's rules. We refer to YR_{ij} as sthe *opponent rule set.*

At the same time they may need to add additional proof trees to their deduction trees in response to questions asked by the other player.

There are six possible statements that can be made in the course of a dialogue:

1. *different_fact*(t, i, j) – i has t as an initial fact and j does not.
2. *different_rule*$(l : A \rightarrow C, i, j)$ – i has $l : A \rightarrow C$ as a rule and j does not.
3. *initial*(t) – t is an initial fact for the Player.
4. $l : a \rightarrow t$ – the player deduced t from the terms in a using the rule labelled l
5. *why*(t) - why do you believe t?
6. *whynot*(t) – why don't you believe t

The first two statements terminate the dialogue.

The starting point for the dialogue is the root term t of T_1 and T_2 which, for the sake of argument, we'll say is provable in T_2 and unprovable in T_1. So t is something one of them has deduced and the other hasn't. We don't worry at this point about how the players have established this initial disagreement.

Definition 6. (Player State) *The* state *of player i at statement k in a dialogue with player j is $S_k^i = \langle D_i, Y_{ij}, F_{ij}, N_{ij}, YR_{ij}\rangle$ where D_i is a deduction (referred to as i's deduction), and $Y_{ij}, F_{ij}, N_{ij}, YR_{ij}$ are sets of terms representing i's opponent belief set, fact, set, disbelief set and rule set respectively.*

Definition 7. (Initial Player State) *The initial state of Player i in a dialogue is ether $\langle \mathcal{D}(F_i, R_i, \{\langle (t, unprovable), \emptyset\rangle\}), \{t\}, \emptyset, \emptyset, \emptyset\rangle$ or $\langle \mathcal{D}(F_i, R_i, \{T_i\}), \emptyset, \emptyset, \{t\}, \emptyset\rangle$ where t is the root term of T_i.*

Definition 8. (Dialogue State) *S_k is the state of the dialogue after the utterance of the kth statement. It consists of the two player states, the last dialogue statement, stmt, and whose turn it is i. $S_k = \langle S_k^1, S_k^2, stmt, i\rangle$*

Definition 9. (Dialogue) *A dialogue is a sequence of dialogue states $S_0, \ldots, S_n$.*

$S_0 = \langle S_0^1, S_0^2, stmt_0, i\rangle$ where $S_0^1 = \langle D_0^1, \{t\}, \emptyset, \emptyset, \emptyset\rangle$ is an initial player state, $S_0^2 = \langle D_0^2, \emptyset, \emptyset, \{t\}, \emptyset\rangle$ is an initial player state, and either $i = 1$ and $stmt_0 = whynot(t)$ (Player 2 started the dialogue by asking Player 1 why they don't believe t and it is now Player 1's turn) or $i = 2$ and $stmt_0 = why(t)$ (Player 1 started the dialogue by asking Player 2 why they believe t and it is now Player 2's turn).

If $S_k = \langle S_k^1, S_k^2, stmt_k, i_k\rangle$ is the state of a dialogue at utterance k then $S_{k+1} = \langle S_{k+1}^1, S_{k+1}^2, stmt_{k+1}, i_{k+1}\rangle$ is a legal next state in the dialogue if the individual parts of the state are as follows:

- (S^i_{k+1} **where** $i = i_k$) *Recall that* $S^i_k = \langle D_i, Y_{ij}, F_{ij}, N_{ij}, YR_{ij}\rangle$
  1. *If* $stmt_k = initial(t)$ *then* $S^i_{k+1} = \langle D_i, Y_{ij} \cup \{t\}, F_{ij} \cup \{t\}, N_{ij}, YR_{ij}\rangle$ *(P_i adds t to Y_{ij} and F_{ij}).*
  2. *If* $stmt_k = l : a \rightarrow t$ *and* $l \in R_i$ *then* $S^i_{k+1} = \langle D_i, Y_{ij} \cup pos(a), F_{ij}, N_{ij} \cup neg(a), YR_{ij} \cup \{l\}\rangle$ *(P_i adds all the positive literals in a to Y_{ij} (these are things the other player believes) and all the negative literals in a to N_{ij} (these are all the things the other player does not believe) and adds l to YR_{ij} – this is a rule the other player has).*
  3. *If* $stmt_k = why(t)$ *and t is provable in one of the deduction trees in D_i then* $S^i_{k+1} = \langle D_i, Y_{ij}, F_{ij}, N_{ij} \cup \{t\}, YR_{ij}\rangle$ *(P_i adds t to N_{ij} (the other player doesn't believe t)).*
  4. *If* $stmt_k = why(t)$, $D_i = \mathcal{D}(F_i, R_i, \mathcal{T})$ *and t is undecided in all the proof trees in $\mathcal{T}$, and there exists a proof tree for F_i and R_i in which t is provable then P_i generates a new proof tree for F_i and R_i, T, with t as its root term and* $S^i_{k+1} = \langle \mathcal{D}(F_i, R_i, \mathcal{T} \cup T), Y_{ij}, F_{ij}, N_{ij} \cup \{t\}, YR_{ij}\rangle$ *(P_i adds t to N_{ij} (the other player doesn't believe t) at the same time P_i generates a deduction as evidence for why it believes t).*
  5. *If* $stmt_k = why(t)$, $D_i = \mathcal{D}(F_i, R_i, \mathcal{T})$ *and* $\langle\{(t, unprovable)\}, \emptyset\rangle$ *is a proof tree for F_i and R_i then* $S^i_{k+1} = \epsilon$ *and the dialogue terminates in an error state (This state should not occur in a legal dialogue – P_i has been asked why it thinks something is true when in fact P_i thinks it is false).*
  6. *If* $stmt_k = whynot(t)$, $D_i = \mathcal{D}(F_i, R_i, \mathcal{T})$ *and t is unprovable in some proof tree in $\mathcal{T}$ then* $S^i_{k+1} = \langle D_i, Y_{ij} \cup \{t\}, F_{ij}, N_{ij}, YR_{ij}\rangle$ *(P_i adds t to Y_{ij} (the other player believes t)).*
  7. *If* $stmt_k = whynot(t)$, $D_i = \mathcal{D}(F_i, R_i, \mathcal{T})$, *$t$ is undecided in all the proof trees in $\mathcal{T}$ or t is provable in some proof tree in $\mathcal{T}$ then* $S^i_{k+1} = \epsilon$ *and the dialogue terminates in an error state (This state should not occur in a legal dialogue – P_i has been asked why it thinks something is false when in fact P_i thinks it is true or has never tried to prove it).*
- (S^j_{k+1} **where** $j \neq i_k$) $S^j_{k+1} = S^j_k$ *(the state of the player who's turn it is not, is unchanged at the next state in the dialogue)*
- ($stmt_{k+1}$) *(Legal moves for player P_{i_k} in their turn in the dialogue). In what follows* $D_{i_k} = \mathcal{D}(F_{i_k}, R_{i_k}, \mathcal{T}_{i_k})$.
  1. $stmt_{k+1} = initial(t)$ *is legal iff* $stmt_k = why(t)$ *and* $t \in F_{i_k}$
  2. $stmt_{k+1} = l : a \rightarrow t$ *is legal iff:*
     - $stmt_k = why(t)$,
     - $t \notin F_{i_k}$,
     - *there exists a proof tree* $\langle N, E\rangle \in \mathcal{T}_{i_k}$ *such that* $(t, l) \in N$
  3. $stmt_{k+1} = whynot(t)$ *is legal iff*
     - $stmt_k \neq why(t)$
     - $\forall l.l \leq k \rightarrow stmt_l \neq whynot(t)$ *(this question has not been asked before)*
     - *there exists a* $T \in \mathcal{T}_{i_k}$ *such that t is provable in T*
     - $t \in N_{i_k j}$ *(Player i_k identifies a term t that it believes and it has established the other doesn't and asks why not).*
  4. $stmt_{k+1} = why(t)$ *is legal if* $stmt_k = whynot(t)$.
  5. $stmt_{k+1} = why(t)$ *is legal if*
     - $stmt_k \neq whynot(t)$,
     - $\forall l.l \leq k \rightarrow stmt_l \neq why(t)$ *(this question has not been asked before)*
     - *there exists a* $T \in \mathcal{T}_{i_k}$ *such that t is unprovable in T and,*
     - $t \in Y_{i_k j}$ *(Player i_k identifies a term t that it does not believe and it has established the other does and asks why).*
  6. $stmt_{k+1} = different\_fact(t, j, i_k)$ *is legal iff* $t \in F_{i_k j}$ *and* $t \notin F_{i_k}$ *(Player i_k has identified that t was an initial fact for Player j but not for Player i_k)*
  7. $stmt_{k+1} = different\_fact(t, i_k, j)$ *is legal iff* $t \in N_{i_k j}$ *and* $t \in F_{i_k}$ *(Player i_k has identified that t was an initial fact for Player i_k but not for Player j)*
  8. $stmt_{k+1} = different\_rule(l, j, i_k)$ *is legal iff* $l \in YR_{i_k j}$ *and there is no rule* $l : A \rightarrow C \in R_{i_k}$ *(Player i_k has identified that l is a rule for Player j but not for Player i_k)*
  9. *no other statements are legal.*
- (i_{k+1}) $i_{k+1} \in \{1, 2\}$ *and* $i_{k+1} \neq i_k$. *(Next player's turn).*

Note that there are some situations where several responses are legal in which case the player may select one. At a later date we may look at heuristics for selecting responses.

4 Proofs

Lemma 1. *If the current state of a dialogue is $\langle S^1_k, S^2_k, initial(t), i\rangle$, $i \neq j$ and $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ then $t \in F_j$ where F_j and t is provable in some $T \in \mathcal{T}_j$.*

Proof. If it is player i's turn then $initial(t)$ was uttered by player j. This could only be uttered if $t \in F_j$.

Morever suppose $\mathcal{D}(F_j, R_j, \mathcal{T}'_j)$ were j's deduction in the previous state. If t is provable in some $T \in \mathcal{T}'_j$ then $\mathcal{T}_j = \mathcal{T}'_j$ and so t is provable in some $T \in \mathcal{T}_j$. If t were undecided in all proof trees in $\mathcal{T}'_j$ then because a proof tree for F_j and R_j exists with t as its root (namely $T = \langle\{(t, initial)\}, \emptyset\rangle$) then $\mathcal{T}_j = \mathcal{T}'_j \cup \{T\}$ so t is provable in some $T \in \mathcal{T}_j$. □

Lemma 2. *If the current state of a dialogue is $\langle S^1_k, S^2_k, l : a \to t, i\rangle$, $i \neq j$ and $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ is j's current deduction then there exists at least one proof tree, $T_j \in \mathcal{T}_j$ such that (t, l) is a node in T_j , $l : a \to t \in R_j$ and for all $t \in pos(a)$, t is provable in T_j; for all $t \in neg(a)$ then t is unprovable T_j; and $l \in R_j$.*

Proof. If it is player i's turn then $l : a \to t$ was uttered by player j. This could only be uttered if $n = \langle t, l\rangle \in N$ for some proof tree $\langle N, E\rangle \in \mathcal{T}_j$. The rest follows from the definition of a proof tree for F_j and R_j. □

Lemma 3. *If the current state of a dialogue is $\langle S^1_{stmt}, S^2_{stmt}, whynot(t), i\rangle$, $i \neq j$ and $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ is j's current deduction then t is provable in some $T \in \mathcal{T}_j$.*

Proof. If it is player i's turn then $whynot(t)$ was uttered by player j. Either this was the opening statement of the dialogue in which case t is the root node of the only proof tree in $\mathcal{T}_j$. Otherwise this could only be uttered if there exists a $T \in \mathcal{T}_j$ such that t is provable in T. □

The equivalent lemma to the above for $why(t)$ is a corollary of the following theoarem.

State updates are correct We need to establish athat for instance $t \in F_{ij}$ iff $t \in F_j$ (i.e., player i only decides player j has t as an initial fact if player j does indeed have j as an initial fact). The same for Y_{ij}, N_{ij} etc. etc. This involves going through the statements j can make that will lead to i updating those states. We also want to show that the error states ϵ never occur.

Theorem 3. *Given two dialogue participants i and j, where $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ is j's deduction and F_{ij} is i's opponent fact set, $F_{ij} \subseteq F_j$. That is i's model of j's initial facts is a subset of j's actual initial facts.*

Proof. We will argue by induction considering each time something is added to F_{ij}.

Base case. In the initial state of the dialogue $F_{ij} = \emptyset$ therefore $F_{ij} \subseteq F_j$.

Step case. A term t is added to F_{ij} when the previous utterance $stmt = initial(t)$. From Lemma 1 we know that in this case $t \in F_j$. Therefore if $F_{ij} \subseteq F_j$ before t is added to it then $F_{ij} \cup \{t\} \subseteq F_j$.

□

Theorem 4. *Given two dialogue participants i and j where $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ is j's deduction and Y_{ij} is i's opponent belief set, then all terms $t \in Y_{ij}$ are provable in some $T \in \mathcal{T}_j$. That is i's model of j's beliefs is a subset of the provable terms in j's deduction.*

Proof. We will argue by induction considering each time something is added to Y_{ij}.

Base case. There are two possible initial states of Y_{ij} given initial deductions $\mathcal{D}(F_i, R_i, \{T_i\})$ and $\mathcal{D}(F_j, R_j, \{T_j\})$ for the two players. Either $Y_{ij} = \{t\}$ where t is a term that is provable in T_j but not in T_i. In this case, since t is provable T_j we are done. Otherwise $Y_{ij} = \emptyset$ and so there are no terms in Y_{ij} and we are also done.

Step case. A term t is added to Y_{ij} following a statment s when:

1. $s = initial(t)$. From Lemma 1 we know that t is provable in some $T \in \mathcal{T}_j$.
2. $s = l : a \to r$ and t is a positive literal in a. From Lemma 2 we know that all terms in $pos(a)$ are provable in some $T \in \mathcal{T}_j$.
3. $s = whynot(t)$. From Lemma 3 we know that in this case t is provable in some $T \in \mathcal{T}_j$.

□

Corollary *If the current state of a dialogue is* $\langle S^1_{stmt}, S^2_{stmt}, why(t), i\rangle$ *and* $\mathcal{D}(F_i, R_i, \mathcal{T}_i)$ *is i's deduction then t is provable in some* $T \in \mathcal{T}_i$. This follows because j can only ask $why(t)$ if either a) it is the initial state of the dialogue in which case t is the root term of the only deduction tree in $\mathcal{T}_i$ by definition of initial player state or b) $t \in Y_{ji}$ where Y_j is j's opponent belief set and we now know that all terms in Y_{ji} are provable in some $T \in \mathcal{T}_i$ or c) in response to $whynot(t)$ which can only be asked if t is provable in some $T \in \mathcal{T}_i$.

Corollary 2 *If the current state of a dialogue is* $\langle S^1_{stmt}, S^2_{stmt}, why(t), i\rangle$ *then* $S^i_{stmt} \neq \epsilon$ *– i.e., the error state never arises in response to a why question.* This follows directly from the previous corollary since t is provable in some proof tree F_i and R_i.

Theorem 5. *Given two dialogue participants i and j, where* $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ *is j's deduction and* N_{ij} *is i's opponent disbelief set, then all terms* $t \in N_{ij}$ *are unprovable in some proof tree* $T \in \mathcal{T}_j$.

Proof. We will argue by induction considering each time something is added to N_{ij}.

Base case. There are two possible initial states of N_{ij} given initial deductions $\mathcal{D}(F_i, R_i, \{T_i\})$ and $\mathcal{D}(F_j, R_j, \{T_j\})$ for the two players. Either $N_{ij} = \{t\}$ where t is a term that is provable in T_i and unprovable T_j and we are done. Otherwise $N_{ij} = \emptyset$.

Step case. A term t is added to N_{ij} following a statement s by j when:

1. $s = l : a \rightarrow r$ and $t \in neg(a)$. From Lemma 2 we know that all terms in $neg(a)$ are unprovable in some $T \in \mathcal{T}_j$.

2. $s = why(t)$. There are three occasions where j can ask $why(t)$. If j has initiated the dialogue then $\mathcal{T}_j = \{\langle (t, unprovable), \emptyset\rangle$ so t is unprovable in the only tree in $\mathcal{T}_j$. Secondly it is uttered in response to $whynot(t)$ asked by i in this case t was already in N_{ij} and so is not added to the set, so we can ignore this case. The last case is when $t \in Y_{ji}$ where Y_{ji} is j's opponent belief set and there exists a $T \in \mathcal{T}_j$ such that t is unprovable in T.

□

Corollary *If the current state of a dialogue is* $\langle S^1_{stmt}, S^2_{stmt}, whynot(t), i\rangle$ *and* $\mathcal{D}(F_i, R_i, \mathcal{T}_i)$ *is i's deduction then there is no proof tree for* F_i *and* R_i *in which t is provable.* This follows because j can only ask $whynot(t)$ if $t \in N_{ji}$ so there is some proof tree for F_i and R_i in which t is unprovable. From this it follows that there is no proof tree for F_i and R_i in which t is provable for any $t \in N_{ji}$ by Theorem 1.

Corollary 2 *If the current state of a dialogue is* $\langle S^1_{stmt}, S^2_{stmt}, whynot(t), i\rangle$ *then* $S^i_{stmt} \neq \epsilon$ *– i.e., the error state never arrises in response to a why not question.* This follows directly from the previous corollary t is not provable in any proof tree F_i and R_i.

Corollary 3 *If the current state of a dialogue is* $\langle S^1_{stmt}, S^2_{stmt}, why(t), i\rangle$ *and* $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ *is j's deduction then t is not provable in any proof tree for* F_j *and* R_j. This follows because j can only ask $why(t)$ if either i asked $whynot(i)$ in which case this is true from our first corollary or if there is a $T \in \mathcal{T}_j$ in which t in unprovable in which case there is no proof tree for F_j and R_j in which t is provable by Theorem 1.

Theorem 6. *Given two dialogue participants i and j, where* $\mathcal{D}(F_j, R_j, \mathcal{T}_j)$ *is j's deduction and* YR_{ij} *is i's opponent rule set, then* $\forall l.l \in R_{ij} \implies \exists a, c.\, l : a \rightarrow c \in R_j$

Proof. We will argue by induction considering each time something is added to YR_{ij}.

Base case. In the initial state of the dialogue $YR_{ij} = \emptyset$ and the theorem is trivially true.

Step case. A rule label l is added to YR_{ij} when the previous statement is $l : a \rightarrow r$. From Lemma 2 we know that there is at least one proof tree, $T \in \mathcal{T}_j$ such that (t, l) is a node in T_j. Since T is a proof tree for F_j and R_j there exists a rule $l : a \rightarrow c \in R_j$ by Theorem 2. □

4.1 Dialogues terminate

Theorem 7. *Given two initial deductions* $D_1 = \mathcal{D}(F_1, R_1, \{T_1\})$ *and* $D_2 = \mathcal{D}(F_2, R_2, \{T_2\})$. *If there are only a finite number of terms that are provable in some proof tree for* F_i *and* R_i, *then any dialogue D starting from* D_1 *and* D_2 *is finite.*

Proof. We observe that the only statements in D that do not terminate the dialogue are $initial(t)$, $l : a \rightarrow c$, $why(t)$ and $whynot(t)$. $inital(t)$, $l : a \rightarrow c$ and $whynot(t)$ can only be made in response to the statement $why(t)$. So it suffices to show that there are only a finite number of terms t about which the statement $why(t)$ can appear in a dialogue and $why(t)$ can only be asked a finite number of times about each of these.

Since we assume that there are only a finite number of terms provable in a proof tree for F_1 and R_1 and only a finite number of terms provable in some proof tree for F_2 and R_2 and we know that $why(t)$ is only asked if t

is provable in some proof three for either F_1 and R_1 or F_2 and R_2 then there are only a finite number of terms t about which $why(t)$ is asked.

If the preceding statement is not $whynot(t)$ then $why(t)$ can only be asked if it has not been asked before, therefore we are only concerned with statements $why(t)$ that following statements $whynot(t)$. However $whynot(t)$ may only be asked once in a dialogue and so $why(t)$ can only respond to this once, even if $why(t)$ has been asked before. Therefore the dialogue terminates (either with a terminating statement or because no legal move can be made). □

Note this means that only some dialogues terminate and this depends on the structure of the facts and rules for each player. However we know that many sets of facts and rules have this property.

4.2 When dialogues terminate, a difference is detected

Now we have to show that a dialogue terminates by someone saying $different\_fact(t, i, j)$ or $different\_rule(l, i, j)$ not just because the dialogue has run out of legal moves.

Theorem 8. *If the kth state in a legal sequence of dialogue states is $\langle S_k^1, S_k^2, s, i\rangle$ and $s \neq different\_fact(t, i, j)$, $s \neq different\_fact(t, j, i)$ and $s \neq different\_rule(l, i, j)$ then there is a legal next dialogue state.*

Proof. Recall that $S_k^i = \langle \mathcal{D}(F_i, R_i, \mathcal{T}_i), Y_{ij}, F_{ij}, N_{ij}, YR_{ij}\rangle$

We establish that in all legal dialogue states one of the following must hold:

1. There exists $t \in N_{ij}$ that is provable in some proof tree, $T \in \mathcal{T}_i$ and $whynot(t)$ has not previously been asked.
2. There exists a $t \in Y_{ij}$ which is unprovable in some proof tree, $T \in \mathcal{T}_i$ and $why(t)$ has not previously been asked.
3. There exists a $t \in F_{ij}$ and $t \notin F_i$
4. There exists some $l \in YR_{ij}$ such that there is no rule $l : a \to c \in R_i$.

We observe that if at least one of the above holds then there is a legal next move.

We consider a proof by cases

1. **Dialogue start.** In this case $Y_{12} = \{t\}$ and $N_{21} = \{t\}$ for some t provable T_2 and unprovable in T_1 (or vice versa). So either 1 or 2 above is true.
2. $s = initial(t)$. This means $t \in F_j$ (for this to be legal utterance). It can only be uttered in response to $why(t)$ said by i so we also know that t is provable in some $T \in \mathcal{T}_j$ (j believes t) (corollary to theorem 4) and that t is not provable in any proof tree for F_i and R_i (corollary 3 to theorem 4). Since t is not provable in any proof tree for F_i and R_i then $t \notin F_i$ (otherwise $\langle\{(t, initial)\}, \emptyset\rangle$ would be a proof tree for F_i and R_i) therefore after updating its state by adding t to F_{ij} there exists a $t \in F_{ij}$ and $t \notin F_i$.
3. $s = l : a \to t$. This can only be uttered in response to $why(t)$ so we know that t is provable in some $T \in \mathcal{T}_j$ and that t is not provable in any proof tree for F_i and R_i. Since t is not provable in any proof tree for F_i and R_i then there is no rule that would allow i to deduce t so either $l : a \to t \notin R_i$ and after l is added to YR_{ij} our fifth property holds. Or there is some term $t \in pos(a)$ where t is not provable in any proof tree for F_i and R_i or some term $t \in neg(a)$ which is provable in some proof tree for F_i and R_i (there is some difference in the antecedants).

   If there is some term $t \in pos(a)$ where t is not provable in some proof tree for F_i and R_i then after adding $pos(a)$ to Y_{ij} there exists a $t \in Y_{ij}$ which is provable in no proof tree for F_j and R_j.

   If there is a $t \in neg(a)$ which is provable in some proof tree for F_i and R_i then after adding $neg(a)$ to N_{ij} there exists $t \in N_{ij}$ that is the root of a proof tree for F_j and R_j and is provable in that tree.
4. $s = why(t)$. In this case $t \in Y_{ji}$, t is unprovable in some $T \in \mathcal{T}_j$ and t is added to N_{ij}. Since all terms in Y_{ji} are provable in some $T \in \mathcal{T}_i$ (Theorem 4) there exists $t \in N_{ij}$ that is provable in some proof tree for F_j and R_j.
5. $s = whynot(t)$. In this case $t \in N_{ji}$, t is provable in some $T \in \mathcal{T}_j$, and t is added to Y_{ij}. Since every term in N_{ji} is unprovable in some $T \in \mathcal{T}_j$ (Theorem 5), there exists a $t \in Y_{ij}$ which is provable in no proof tree for F_j and R_j (by Theorem 1).

□

Theorem 9. *If a dialogue terminates the last statement is:* $different\_fact(t, i, j)$ *or* $different\_fact(t, j, i)$ *or* $different\_rule(l, i, j)$

Proof. This follows because we have shown that there is always a legal next move except in these cases. □

References

[Dennis and Oren, 2021] Dennis, L. A. and Oren, N. (2021). Explaining BDI agent behaviour through dialogue. In *20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2021)*.