Mind-reading for machines: Inferring and predicting a user's beliefs

Richard Booth
Mahasarakham University
Faculty of Informatics
Mahasarakham 44150, Thailand
richard.b@msu.ac.th

Alexander Nittka
University of Leipzig
Dept. of Computer Science
Leipzig 04109, Germany
nittka@informatik.uni-leipzig.de

Abstract

Imagine the following simple dialogue between an "intelligent" machine and a user:

User: Is Miss Thailand very beautiful? (B?)
Machine: Yes. (B)
User: Then I believe she is Miss Universe. (U)

All the machine explicitly knows from this is that *(i)* at the start of the dialogue, the user did not know whether B was true, and *(ii)* after receiving statement B the user believes U. But did the user believe U also before receiving B? Does he believe B itself after receiving it? And will he still believe B if the machine then tells him, "actually, Miss Puerto Rico is Miss Universe" (i.e., U is false!). Helping the machine find answers to questions like these is the subject of this paper. Our solution depends on the machine building a plausible model of the user and how it forms beliefs, which "best explains" the observed dialogue. Our results will be applicable in the areas of automated dialogue (e.g., chatbots) and user modelling.[1]

1 Introduction

Imagine the following simple dialogue between an "intelligent" machine and a user:

User: Is Miss Thailand very beautiful? (B?)
Machine: Yes. (B)
User: Then I believe she is Miss Universe. (U)

All the machine can explicitly reasonably assume from this about the user's beliefs is that *(i)*

[1] This paper is a more informal exposition of ideas presented in [2, 3].

at the start of the dialogue, the user did not know whether B was true (otherwise he would not have asked), and *(ii)* after receiving statement B the user believes U. But did the user believe U also *before* receiving B? Does he believe B itself after receiving it? And suppose the dialogue continues with:

Machine: Actually, Miss Puerto Rico is Miss Universe. (i.e., U is false!)

Will the user still believe B after receiving this information? Helping the machine find answers to questions like these is the subject of this paper.

Our solution depends on the machine building a plausible model of the user and how it forms beliefs, which "best explains" the observed dialogue. The plan of the paper is as follows. In Section 2 we describe the way we take the machine to *internally represent* a dialogue. The method of representation is based on *propositional logic*, a mathematically precise formalism which is widely used in the field of Knowledge Representation. In Section 3 we describe the model of the user which the machine should use and define what it means for one such model to *explain* a dialogue. Then in Section 4 we investigate what it means for a given model to be a *best explanation* for a given dialogue. Throughout the paper we will illustrate our ideas on the above Miss Thailand example, leading to the best-explaining model for this particular dialogue. We conclude and mention ideas for future research in Section 5.

2 Representing the dialogue

Automated modelling of natural language dialogues such as the above Miss Thailand dialogue requires *two* separate phases:

Phase 1 the use of some *natural language processing (NLP)* module to transform the natural language statements into some representation suitable for manipulation by the machine.

Phase 2 the use of some *reasoning mechanism* by which the machine can carry out these manipulations.

Although both phases are equally important, in this paper we will focus *entirely* on Phase 2. We will see that already just in this phase some interesting issues emerge. Furthermore we shall assume the suitable representation mentioned in Phase 1 is the language of *propositional logic*. In propositional logic, statements are represented as sentences which are built up from a set of *propositional variables* (sometimes also known as *atomic sentences*) using the logical connectives $\&$, $\vee$, $\neg$ and $\rightarrow$ standing for "and", "or", "not" and "if ... then ..." (or "implies") respectively. For example for the purposes of modelling the Miss Thailand example we might choose propositional variables `v_beautiful` and `Miss_Universe` to stand for the statements "Miss Thailand is very beautiful" and "Miss Thailand is Miss Universe" respectively. Using the connectives we can then express more complex sentences such as:

$\texttt{v\_beautiful} \,\&\, \texttt{Miss\_Universe}$
"Miss Thailand is very beautiful *and* she is Miss Universe"

$\neg\texttt{v\_beautiful} \rightarrow \neg\texttt{Miss\_Universe}$
"*If* Miss Thailand is *not* very beautiful *then* she is *not* Miss Universe"

$\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}$
"*If* Miss Thailand is very beautiful *then* she is Miss Universe"

$\neg\texttt{v\_beautiful} \vee \texttt{Miss\_Universe}$
"Either Miss Thailand is *not* very beautiful *or* she is Miss Universe (or both)"

(For an introduction to propositional – also sometimes called *sentential* – logic, see, e.g., [4, Chapter 1].) We propose that statements provided by both the machine and the user be modelled internally by the machine as sentences in propositional logic. We do not concern ourselves with the precise method of *how* the natural language statements get converted into this representation, we just assume this process of conversion has already taken place.

As well as a way to internally represent the statements exchanged in the dialogue, the machine needs some way, or *data structure*, to model the dialogue as a whole. We propose the machine stores the dialogue as a series of *stages*, or *rounds*, in each of which the user and machine each take it in turn to provide an input. Each of these individual stages is assumed to follow a rigid pattern in which *first* the machine makes some statement to the user, and *then* the user provides some *response* to the machine's input. From the user's response, we assume the machine is always able to work out a set of propositional sentences which the user is taken to *believe* at that moment in time, **and also** a set of sentences which the user is taken *not* to believe at that moment. Formally, each stage of dialogue consists of a triple (a_i, B_i, D_i), where

- a_i is a (propositional) sentence representing the machine's input to the user.
- B_i and D_i are both (finite) *sets* of sentences which collectively indicate the user's *response* to receiving a_i:
  - B_i is a set of sentences which the user is taken to *believe* in response to a_i.
  - D_i is a set of sentences which the user is taken *not* to believe in response to a_i.

The D_i can arise quite naturally in the context of a dialogue. For example if, following being told a_i, the user responds with the question "Is b true or not?", then (assuming the user is being sincere) the machine may assume the user believes neither b nor $\neg b$, and so both b and $\neg b$ would be elements of D_i.

Our insistence that the machine always takes the first turn in a round of dialogue might seem too restrictive. After all, as the Miss Thailand example shows, sometimes it is the user who *starts* the dialogue. However, as we will soon see, there is a way of working around this restriction.

We assume the machine then stores the full dialogue as a *sequence* of such stages

$$d = \langle (a_1, B_1, D_1), (a_2, B_2, D_2), \ldots, (a_n, B_n, D_n) \rangle,$$

where (a_1, B_1, D_1) is the first round, (a_2, B_2, D_2) is the second round, and so on up to the last round (a_n, B_n, D_n).

Example 2.1. *Let's take a look at how the Miss Thailand example might be modelled in this way. We can use two dialogue rounds. First note the dialogue as it stands doesn't fit into our rigid turn-taking pattern of machine-user-machine-user-..., since the first contribution is by the user rather than the machine. However we can easily make it fit by assuming an imaginary first "dummy" turn by the machine in which the machine provides the user with an empty input which we denote by* $\top$. *Formally speaking, in propositional logic* $\top$ *denotes a sentence which is* always *true (sometimes called a "tautology"), and therefore provides the user with no useful information, since the user already knows it is true. The user responds to this empty input by asking if Miss Thailand is very beautiful. Thus the set of sentences the user* doesn't *believe at this stage can be taken to include* $\texttt{v\_beautiful}$ *and* $\neg\texttt{v\_beautiful}$. *The user's response does not give the machine any information about what he* does *believe. Thus the first stage can be modelled as*

$$(\top, \emptyset, \{\texttt{v\_beautiful}, \neg\texttt{v\_beautiful}\}).$$

Here, $\emptyset$ *denotes the "empty set", i.e., the set containing no elements. In the next stage, the machine provides the user with the information that Miss Thailand is very beautiful, and the user responds by asserting that Miss Thailand is Miss Universe. Thus the second stage can be modelled by*

$$(\texttt{v\_beautiful}, \{\texttt{Miss\_Universe}\}, \emptyset).$$

This is the end of the dialogue, which we denote by d_{Miss}.

$$d_{Miss} = \langle(\top, \emptyset, \{\texttt{v\_beautiful}, \neg\texttt{v\_beautiful}\}), (\texttt{v\_beautiful}, \{\texttt{Miss\_Universe}\}, \emptyset)\rangle.$$

Now, speaking more generally, the problems we are interested in may be stated as follows: assume the machine is given some piece of dialogue

$$d = \langle(a_1, B_1, D_1), (a_2, B_2, D_2), \ldots, (a_n, B_n, D_n)\rangle.$$

Then the machine must answer the following questions:

- What did the user believe before the *start* of the dialogue, i.e., before receiving a_1?
- For each $i = 1, \ldots, n$, what *apart from* the sentences in B_i, did he believe after receiving a_i?
- What will he believe after receiving a given further input a_{n+1}?

We make two assumptions about this setting: *(i)* $(a_1, \ldots, a_n)$ represent *all* the inputs the user receives between during the course of the dialogue, i.e., the user does not receive any other information during this time from any other source. *(ii)* the B_i and the D_i are *correct* (but possibly *incomplete*) descriptions of the user's beliefs and non-beliefs, i.e., following each input from the machine a_i, the user really *does* believe all the sentences in B_i and really does *not*

believe any sentence in D_i. In other words the user is being sincere with his responses, and is not lying or trying to mislead the machine.

How can the machine answer these questions? Our idea is that the machine needs to somehow "get inside the mind" of the user, and form some sort of picture, or *model*, of how the user forms beliefs and updates his beliefs when he receives new information during a dialogue. The machine can then use this model to *explain* the user's responses during the dialogue, and to make extra inferences and predictions about the user's beliefs. The question is: what should this model look like? What should be the *ingredients* of such a model? We now look at one possibility for describing a user's *belief state*, originally described in [1].

3 Modelling the user

The model of the user's belief state needs to be simple yet have a touch of realism, or plausibility. The idea behind the model we propose can be said to come from two basic observations about how we, as human beings, form beliefs:

(a) our beliefs depend on the information we have received over time, and

(b) we often have some beliefs which we believe so strongly that we will never give them up, for example "The capital city of England is London". We will call these *core* beliefs.

Consider again the user. For *(a)* we let $\mathbf{a}$ denote a sequence $(b_1, b_2, \ldots, b_m)$ of (propositional) sentences standing for the sequence of inputs the user has received so far (with b_1 being the first input received, and b_m the last), while for *(b)* we let ▲ be a (propositional) sentence standing for the core beliefs of the user. Then the belief state of the user should contain some record of $\mathbf{a}$ and ▲. In fact we take $\mathbf{a}$ and ▲ together *to be* the belief state of the user, which we will denote by $[\mathbf{a} \mid ▲]$.

Example 3.1. *For a simple example suppose the user has received during his life three items of information: p followed by q and then $\neg q$, where p and q denote some given (distinct) propositional variables. Then $\mathbf{a} = (p, q, \neg q)$. Suppose the user has r (another propositional variable) as his only core belief, i.e., ▲ $= r$. Then the user's belief state is $[(p, q, \neg q) \mid r]$.*

We're not yet finished with our model of the user, we still need to describe *(i)* given the user is in state $[\mathbf{a} \mid ▲]$ what are his actual *beliefs* in this state (i.e., the sentences the user actually uses when making decisions)? We also need to describe *(ii)* how the user *changes* his belief state whenever he receives new information. Turning first to *(i)*, an obvious first answer would be to say the user believes all the information he has received *plus* the core beliefs. However this might not always be possible, since it is very likely that the user, during the course of his life, has received information which is *contradictory*. Intuitively, this means the user has received two or more items of information which *cannot all be true at the same time*. The most typical example would

be if the user has received an input sentence a, and then at some further point in time receives the opposite information $\neg a$. In this case, if the user believes simultaneously *all* information he has ever received, then, to use the correct technical term, his beliefs become *inconsistent*. And it is a basic instinct of the user to *avoid* having inconsistent beliefs. Thus in this situation the user needs to employ some method to choose between contradictory information. The machine should assume the user is using a very simple mechanism to do this. One such mechanism is that, if there is an inconsistency, then the user gives *priority* to more *recent* information received. On top of this, the beliefs must include ▲, since these are the core beliefs. So ▲ receives the *highest* priority of all the sentences appearing in the user's belief state.

One method which takes all this into account can be roughly[2] described as follows. For the moment let us assume $\mathbf{a} = (b_1, b_2, \ldots, b_m)$. The user calculates his beliefs in stages, starting with ▲. First the user looks at sentence b_m, i.e., the most recent sentence received. If the sentence $b_m \,\&\, ▲$ is consistent, then it's OK for the user to believe b_m. So he updates his beliefs from ▲ to $b_m \,\&\, ▲$ and then looks at the next sentence along in the sequence b_{m-1}. If, however $b_m \,\&\, ▲$ is *in*consistent then the user should not believe b_m. In this case the user keeps his beliefs as just ▲ and then moves along to consider b_{m-1}. Now if b_{m-1} is consistent with the sentence collected so far ($b_m \,\&\, ▲$ in the first case, just ▲ in the second) then it is added, and then the next sentence b_{m-2} is considered. If it is *in*consistent with the sentence collected so far, then it is not added, and then the next sentence is considered. The process continues like this, working backwards through $\mathbf{a}$, until all the sentences in $\mathbf{a}$ have been considered. We denote the beliefs of the user in state $[\mathbf{a} \mid ▲]$ by $Bel([\mathbf{a} \mid ▲])$.

Example 3.2. *Assume the user is in the belief state from Example 3.1, i.e.,* $[(p, q, \neg q) \mid r]$. *Then the user believes* r. *Since* $\neg q \,\&\, r$ *is consistent we may add* $\neg q$. *But at the next stage, since adding* q *to* $\neg q \,\&\, r$ *would give an inconsistency,* q *is not added and we move on to consider the remaining sentence* p. *Adding* p *to* $\neg q \,\&\, r$ *does not give an inconsistency, so it may be added, and so we finish with* $Bel([(p, q, \neg q) \mid r]) = p \,\&\, \neg q \,\&\, r$.

Note that $Bel([\mathbf{a} \mid ▲])$ is always a *single* propositional sentence, in fact a collection of propositional sentences connected by "&". However, when using a single sentence to represent a user's beliefs, we are implicitly assuming that the user believes not just that sentence, but in fact all *logical consequences* of that sentence. For instance, in the above example the full set of statements which the user believes contains not just $p \,\&\, \neg q \,\&\, r$, but also sentences like p, $\neg q$, r, $p \vee q$, etc, as well as the tautology $\top$ (which is a logical consequence of *every* sentence). Thus in the rest of the paper we will sometimes talk about $Bel([\mathbf{a} \mid ▲])$ *as though* it was a *set* of sentences. In these cases the understanding is that $Bel([\mathbf{a} \mid ▲])$ is the set containing all logical consequences of the sentence being used to represent it.

This, then, describes how the user calculates his beliefs in any given belief state. But how does he *change* his belief state when he receives new information. Given the above model of the user's belief state, the procedure is easy. Given his current belief state is $[\mathbf{a} \mid ▲]$ and he receives new input sentence b, the user just adds b to the right-hand end of $\mathbf{a}$, i.e., the user's new belief state becomes $[\mathbf{a} \cdot b \mid ▲]$, where "$\cdot$" is just the "append" operator. Then his new belief set can be worked out again using the same procedure outlined above.

[2]For the more formal description see [2, 3].

Example 3.3. *Again suppose the user is in the belief state* $[(p, q, \neg q) \mid r]$.
(i) *Suppose the user receives new information* $\neg q \rightarrow \neg r$*, so his new belief state is*

$$[(p, q, \neg q, \neg q \rightarrow \neg r) \mid r].$$

When calculating his new set of beliefs, the user starts with r*. Adding* $\neg q \rightarrow \neg r$ *to* r *does not lead to inconsistency, so it may be added. But adding* $\neg q$ *to* $(\neg q \rightarrow \neg r)$ & r does *lead to inconsistency (since* $\neg q$ *together with* $\neg q \rightarrow \neg r$ *logically implies* $\neg r$*, which contradicts the core belief* r*). So whereas* $\neg q$ *was a part of the user's belief set before, the new information causes the user to* give up *this belief. The remaining two sentences in the sequence may both be added to* $(\neg q \rightarrow \neg r)$ & r *without causing inconsistency, so the user's new belief set is* p & q & $(\neg q \rightarrow \neg r)$ & r*. Since* $\neg q \rightarrow \neg r$ *is actually a* logical consequence *of the other three sentences appearing here it may be removed without changing the meaning of the sentence, i.e., the user's belief set here can be equivalently described as just* p & q & r*. Throughout the rest of the paper, we will often repeat this trick of removing redundant sentences in order to simplify the description of the user's beliefs. In particular note that, since the sentence* $\top$ *is a logical consequence of* any *sentence, we are* always *free to remove* $\top$.
(ii) *Suppose instead the user receives new information* $\neg r$*, so his new belief state becomes* $[(p, q, \neg q, \neg r) \mid r]$*. When calculating his new beliefs, the user starts as always with his core beliefs* r*. But then since adding* $\neg r$ *would give inconsistency, the user does* not *include* $\neg r$ *in his beliefs. In fact his beliefs in this new state stay the same as the beliefs in the original state, i.e.,* p & $\neg q$ & r*. This example shows there are times when the user does not believe new information given to him, namely when that information contradicts his core beliefs.*

This completes our description of how the user forms and updates his beliefs. **We now instruct the machine to assume the user works like this.** Then, given a dialogue

$$d = \langle (a_1, B_1, D_1), (a_2, B_2, D_2), \ldots, (a_n, B_n, D_n) \rangle,$$

the machine knows what the user's belief state looks like at each step, namely

$$[\mathbf{a} \cdot (a_1, \ldots, a_i) \mid \blacktriangle],$$

and his belief set is $Bel([\mathbf{a} \cdot (a_1, \ldots, a_i), \blacktriangle])$, where $[\mathbf{a} \mid \blacktriangle]$ is the user's *initial* (i.e., before receiving a_1) belief state. The only problem is the machine does *not* know what $[\mathbf{a} \mid \blacktriangle]$ is. The dialogue d gives the machine the following information about the user's beliefs:

$$\begin{array}{l} \text{for all } i \text{ such that } 1 \leq i \leq n: \\ B_i \subseteq Bel([\mathbf{a} \cdot (a_1, \ldots, a_i) \mid \blacktriangle]) \text{ and} \\ D_i \cap Bel([\mathbf{a} \cdot (a_1, \ldots, a_i) \mid \blacktriangle]) = \emptyset. \end{array} \quad (1)$$

(Here "$\subseteq$" denotes the subset relation between sets, and "$\cap$" denotes the operation of taking the *intersection* of two sets: $X \cap Y$ is the set containing all elements which belong to both X *and* Y.)

We make the following definition:

Definition 3.4. *We say* $[\mathbf{a} \mid \blacktriangle]$ explains *a dialogue* d *(or is an* explanation for d*) if (and only if) equation (1) above holds.*[3]

For a given dialogue d and belief state $S = [\mathbf{a} \mid \blacktriangle]$, for each $i = 0, 1, \ldots, n$ let us write Bel_i^S as shorthand for $Bel([\mathbf{a} \cdot (a_1, \ldots, a_i) \mid \blacktriangle])$ (so $Bel_0^S = Bel([\mathbf{a} \mid \blacktriangle])$ is the user's initial beliefs at the start of the dialogue). Then we will call the sequence

$$(Bel_0^S, Bel_1^S, \ldots, Bel_n^S)$$

the *belief trace of* S *through* d. It gives the evolution of the user's beliefs through the dialogue d *if* the user's initial belief state is assumed to be $[\mathbf{a} \mid \blacktriangle]$.

3.1 The Miss Thailand example

Let's return to our Miss Thailand dialogue

$$d_{Miss} = \langle (\top, \emptyset, \{\texttt{v\_beautiful}, \neg\texttt{v\_beautiful}\}), (\texttt{v\_beautiful}, \{\texttt{Miss\_Universe}\}, \emptyset)\rangle.$$

We will now give one non-explanation and four different explanations for d_{Miss}. According to Definition 3.4, in order for a given $S = [\mathbf{a} \mid \blacktriangle]$ to *explain* d_{Miss} we require

$$(1)\ \texttt{v\_beautiful} \notin Bel_1^S \quad (2)\ \neg\texttt{v\_beautiful} \notin Bel_1^S \quad (3)\ \texttt{Miss\_Universe} \in Bel_2^S$$

(Here, "$\in$" denotes set-membership, i.e., means "is an element of", "$\notin$" means "is not an element of".)
(i). Consider the belief state

$$S_1 = [(\texttt{v\_beautiful}\ \&\ \texttt{Miss\_Universe}) \mid \top)],$$

i.e., $\mathbf{a}$ is here taken to be the sequence whose only element is $\texttt{v\_beautiful}\ \&\ \texttt{Miss\_Universe}$, and the user has no core beliefs.[4] The belief trace of S_1 through d_{Miss} is

$$\begin{array}{l}(\texttt{v\_beautiful}\ \&\ \texttt{Miss\_Universe},\\ \ \texttt{v\_beautiful}\ \&\ \texttt{Miss\_Universe},\\ \ \texttt{v\_beautiful}\ \&\ \texttt{Miss\_Universe}).\end{array}$$

(Note that, in *all* the examples of this subsection, the first two elements of the belief trace will be the same since the first input the user receives from the machine is the "empty" input $\top$ which has no effect on the user's beliefs.) This belief state does *not* explain d_{Miss}, because $\texttt{v\_beautiful} \in Bel_1^{S_1}$, violating (1) above. The rest of our examples are all explanations for d_{Miss}.

[3] In [3] the additional condition "$\blacktriangle$ is consistent" was required in order for $[\mathbf{a} \mid \blacktriangle]$ to qualify as an explanation. However this is not needed for the purpose of this paper.

[4] At least no *interesting* core beliefs. Tautologies, e.g., $\texttt{Miss\_Universe} \rightarrow \texttt{Miss\_Universe}$ are *always* included as core beliefs.

(ii). Consider

$$S_2 = [() \mid \texttt{Miss\_Universe}],$$

i.e., the user did not receive *any* previous inputs, but has `Miss_Universe` as a core belief. Thus the user has a blind conviction that Miss Thailand is Miss Universe (whether she is very beautiful or not!). The belief trace is

$$(\texttt{Miss\_Universe},\\ \texttt{Miss\_Universe},\\ \texttt{v\_beautiful} \,\&\, \texttt{Miss\_Universe}).$$

(iii). Consider

$$S_3 = [(\texttt{Miss\_Universe}) \mid \top],$$

i.e, the user again starts off believing Miss Thailand is Miss Universe, but this time it is not a core belief, i.e., the user is prepared to give up this belief if evidence arrives saying otherwise. The belief trace of S_3 through d_{Miss} is the same as that of S_2:

$$(\texttt{Miss\_Universe},\\ \texttt{Miss\_Universe},\\ \texttt{v\_beautiful} \,\&\, \texttt{Miss\_Universe}).$$

(iv). Consider

$$S_4 = [(\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}, (\texttt{v\_beautiful} \,\&\, \neg\texttt{Miss\_Universe}) \rightarrow \texttt{fix}) \mid \top],$$

where `fix` is a variable standing for the statement "The Miss Universe contest is a fix!". So the user has learned two prior facts: that Miss Thailand is Miss Universe *if* she is very beautiful, and that *if* Miss Thailand is very beautiful but is *not* Miss Universe, then the contest must be a fix. We have for the belief trace

$$(\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe},\\ \texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe},\\ \texttt{v\_beautiful} \,\&\, \texttt{Miss\_Universe}),$$

i.e., at the start of the dialogue (and following the empty first input $\top$) the user does *not* believe Miss Thailand is Miss Universe, only that she is Miss Universe *if* she is very beautiful. Following the confirmation by the machine that she is indeed very beautiful, the user believes that what the machine says is true, and then naturally also believes she is Miss Universe.

(v). Consider the previous example but without the second input concerning the fix, i.e.,

$$S_5 = [(\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}) \mid \top].$$

We get the exact same belief trace as in *(iv)* above. As we will soon see, the difference between explanations S_4 and S_5 only reveals itself when we consider their consequences for *predicting* what the user will believe following *future* inputs.

As the Miss Thailand example shows, it is possible for dialogues in general to have more than one possible explanation. The reader might be wondering whether there are dialogues which

have *no* explanation. The following simple example shows that such unexplainable dialogues do indeed exist.

$$\langle(\top, \{a\}, \{a\})\rangle,$$

where a is any sentence. This dialogue states that the user both believes a and *doesn't* believe a at the same time, which is clearly impossible. Another more subtle example of an unexplainable observation is

$$\langle(p, \emptyset, \emptyset), (q, \emptyset, \{p, \neg p\})\rangle,$$

where p and q are distinct propositional variables. In other words the observation is saying that after receiving p followed by q, the user is undecided whether to believe p or $\neg p$. In fact it is a property of our user model of Section 3 that whenever the user received a new input a, then after *every subsequent input* he either believes a or he believes $\neg a$ (although he might oscillate between the two). So, assuming the *particular* model of an agent's belief state being used in this paper, the above dialogue cannot be explained. Of course that is not to say that it cannot be explained using some *other*, perhaps more sophisticated model. In the rest of this paper we always assume the given dialogue d is explainable.

Now, returning to the general problem, *if* the machine could find *some* explanation $S = [\mathbf{a} \mid \blacktriangle]$ for d, then it would be able to answer all the questions from page 4: before the dialogue began the user believed Bel_0^S, after receiving each a_i the user believed Bel_i^S, while given any further input a_{n+1} the user will be predicted to believe $Bel([\mathbf{a} \cdot (a_1, \ldots, a_n, a_{n+1}) \mid \blacktriangle])$. But, as we've just seen, several explanations might exist. How does the machine choose between them? The short answer is that it should choose the *best* one. But what does "best" mean here? What makes one explanation "better than" another? The guiding intuition that we follow in this work is that, when making inferences and predictions about the user's beliefs, the machine should try and stick to forming conclusions which are *justified on the basis of the given dialogue alone*. We consider a good explanation to be one which goes as little as possible *beyond* the information given explicitly by the dialogue. To put it another way, the machine should be *cautious* in the inferences it makes, it should not be too quick to jump to *bold* conclusions about the user. (This same philosophy is behind the familiar *maximum entropy* approach to probabilistic inference [6].) However this is still a rather vague description. How can we formalise this precisely? In the next section we look some ways in which we can do this.

4 Finding the best explanation

When given a collection X of possible options over which we want to make a choice (in our case the collection of possible explanations for the given dialogue d), a mathematically standard way to proceed is to formally define a *preference relation over* X, that is, define some binary relation $\preceq$ among the options in X with the intuition that, given any two options x and y, $x \preceq y$ holds precisely when "y is at least as good (or preferred) as x". A *best* (or maximally preferred) option in X is then any option which is at least as good as all other options, i.e., any option z such that $x \preceq z$ *for all other options* x. If $x \preceq y$ but $y \not\preceq x$ then we say y is *strictly better than* x and write this as $x \prec y$, while if both $x \preceq y$ *and* $y \preceq x$, then this means x and y are considered *equally good*. We write this as $x \sim y$. We will now define a series of *three* different possible

preference relations over the collection of explanations for d, and we will illustrate each one on the explanations for the Miss Thailand example from Section 3.1.

4.1 First preference relation $\preceq_1$

As we said earlier, we want to prefer explanations which make the fewest assumptions about the user. This applies in particular to the *core beliefs* of the user. We prefer explanations which lead the machine to infer the user has as *few* core beliefs as possible. Indeed the user's set of (non-tautologous) core beliefs should be take to be empty if possible. This leads us to define the following preference relation, given any two explanations $[\mathbf{a}_1 \mid \blacktriangle_1]$ and $[\mathbf{a}_2 \mid \blacktriangle_2]$ for d:

$$[\mathbf{a}_1 \mid \blacktriangle_1] \preceq_1 [\mathbf{a}_2 \mid \blacktriangle_2] \;\Leftrightarrow\; \blacktriangle_2 \subseteq \blacktriangle_1.$$

We propose that the best explanation for d should *at the very least* be a best explanation according to $\preceq_1$.

Example 4.1. Looking at the explanations from the Miss Thailand example, we have the set core beliefs of explanation S_2 is `Miss_Universe`, while the core of explanations S_3, S_4 and S_5 is empty. Thus we have $S_2 \prec_1 S_i$ for each $i = 3, 4, 5$ and $S_i \sim_1 S_j$ for each $i, j = 3, 4, 5$ such that $i \neq j$. In other words S_3, S_4 and S_5 are all equally good and are all strictly better than S_2 according to $\preceq_1$. This means S_2 cannot be regarded as a best explanation. Furthermore S_3, S_4 and S_5 clearly can't be bettered according to $\preceq_1$ – they are all best explanations according to $\preceq_1$ – so we need some further criteria to sort these.

Note the best explanations (according to $\preceq_1$) in the Miss Thailand example all have the empty core. This is because there is nothing in the dialogue d_{Miss} which *forces* the user to have *any* core beliefs. In other words core beliefs are not necessary to explain this dialogue. For an example of a dialogue which is explainable but nevertheless only assuming a *non-empty* set of core beliefs on the part of the user take $\langle (p, \emptyset, \emptyset), (q, \emptyset, \{p\}) \rangle$ (where, again, p and q are distinct propositional variables), i.e., after receiving p followed by q, the user does *not* believe p. In fact *any* explanation for this observation must include *at least* $q \rightarrow \neg p$ as a core belief.

4.2 Second preference relation $\preceq_2$

As we see from the above example, $\preceq_1$ still leaves a lot of choice in the search for a best explanation. It can happen that there are still quite a large number of explanations which are "best" according to $\preceq_1$. How can we narrow down the possibilities a bit further? We now bring in a second preference relation $\preceq_2$ which will be used to do this. This relation will be defined by comparing the *belief traces* through d of two given explanations. The basic idea is that a belief should be ascribed to the user as *late* in the dialogue as possible. In particular, given two explanations S and T, we should prefer that which leads the machine to infer the user has *fewer* beliefs at the *commencement* of the dialogue. In case they both give the *same* beliefs

at this stage, i.e., $Bel_0^S = Bel_0^T$, we prefer that explanation which leads the machine to infer the user has fewer beliefs after the *first* input a_1. In case also $Bel_1^S = Bel_1^T$ we then prefer that explanation which gives fewer beliefs after the next input a_2, and so on. All this can be expressed by defining $\preceq_2$ as follows:

$$[\mathbf{a}_1 \mid \blacktriangle_1] \preceq_2 [\mathbf{a}_2 \mid \blacktriangle_2] \;\Leftrightarrow\; Bel_k^{[\mathbf{a}_2|\blacktriangle_2]} \subseteq Bel_k^{[\mathbf{a}_1|\blacktriangle_1]}$$
$$\text{where } k \text{ is least such that } Bel_k^{[\mathbf{a}_2|\blacktriangle_2]} \neq Bel_k^{[\mathbf{a}_1|\blacktriangle_1]}$$

Example 4.2. Having already discarded S_2 as a candidate for best explanation of d_{Miss} on the basis of $\preceq_1$, let us compare the remaining explanations S_3, S_4 and S_5 with respect to $\preceq_2$. Looking at the initial beliefs, we have $Bel_0^{S_3} = \texttt{Miss\_Universe}$ and $Bel_0^{S_4} = Bel_0^{S_5} = \texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}$. Since $\texttt{Miss\_Universe}$ logically implies $(\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe})$ we have $Bel_0^{S_4} \subseteq Bel_0^{S_3}$ and $Bel_0^{S_5} \subseteq Bel_0^{S_3}$. However $Bel_0^{S_3}$ contains a belief, namely $\texttt{Miss\_Universe}$, which is *not* in $Bel_0^{S_4}$ (and $Bel_0^{S_5}$). Hence $Bel_0^{S_3} \not\subseteq Bel_0^{S_4}$ and $Bel_0^{S_3} \not\subseteq Bel_0^{S_4}$. This means we have both $S_3 \prec_2 S_4$ and $S_3 \prec_2 S_5$, i.e., S_4 and S_5 are *strictly* preferred to S_3 according to $\preceq_2$. S_3 cannot be considered a best explanation according to $\preceq_2$ since it unjustifiably leads the machine to infer an initial belief on the user's part, namely $\texttt{Miss\_Universe}$. Since the belief traces of S_4 and S_5 are the same, these two explanations clearly cannot be distinguished on the basis of $\preceq_2$. Some further criterion is necessary in order to do this.

4.3 Third preference relation $\preceq_3$

A third possibility to compare explanations is to look at their consequences for *predicting* what the user would believe following a *further* input a_{n+1} from the machine. Recall that we want the machine to be *cautious* in the predictions it makes. This means we should prefer those explanations which always lead to *fewer* beliefs being predicted. This is the precisely the idea behind our third preference relation $\preceq_3$:

$$[\mathbf{a}_1 \mid \blacktriangle_1] \preceq_3 [\mathbf{a}_2 \mid \blacktriangle_2] \;\Leftrightarrow\; \text{for } \textit{all} \text{ possible further inputs } a_{n+1},$$
$$Bel([\mathbf{a}_2 \cdot (a_1, \ldots, a_n, a_{n+1}) \mid \blacktriangle_2]) \subseteq$$
$$Bel([\mathbf{a}_1 \cdot (a_1, \ldots, a_n, a_{n+1}) \mid \blacktriangle_1]).$$

Example 4.3. In the Miss Thailand example we have found that we cannot split explanations S_4 and S_5 according to $\preceq_1$ and $\preceq_2$. Let's see if we can separate them according to $\preceq_3$. First, to make things easier to read, let us write

$$\mathbf{a}_4 = (\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}, (\texttt{v\_beautiful} \;\&\; \neg\texttt{Miss\_Universe}) \rightarrow \texttt{fix})$$

and

$$\mathbf{a}_5 = (\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}).$$

So $S_4 = [\mathbf{a}_4 \mid \top]$ and $S_5 = [\mathbf{a}_5 \mid \top]$. Now suppose the dialogue continues with the machine telling the user that Miss Thailand is *not* Miss Universe, i.e., $\neg\texttt{Miss\_Universe}$. We have

$$Bel(\mathbf{a}_4 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{Miss\_Universe}) \mid \top) =$$
$$\texttt{v\_beautiful} \;\&\; \neg\texttt{Miss\_Universe} \;\&\; \texttt{fix}$$

i.e., explanation S_4 predicts that, after being informed that Miss Thailand is not Miss Universe, the user will believe the Miss Universe contest is fixed! Meanwhile we have

$$Bel(\mathbf{a}_5 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{Miss\_Universe}) \mid \top) = \texttt{v\_beautiful} \,\&\, \neg\texttt{Miss\_Universe},$$

i.e., explanation S_5 predicts that after this extra information, the user will simply believe that Miss Thailand is still very beautiful, but is not Miss Universe. According to this explanation, the user will *not* believe the contest is fixed. Thus we see that

$$Bel(\mathbf{a}_5 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{Miss\_Universe}) \mid \top) \subseteq Bel(\mathbf{a}_4 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{Miss\_Universe}) \mid \top),$$

but

$$Bel(\mathbf{a}_4 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{Miss\_Universe}) \mid \top) \not\subseteq Bel(\mathbf{a}_5 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{Miss\_Universe}) \mid \top).$$

Now in fact it can be proved that, for *any* possible further input a_3 (not just $\neg\texttt{Miss\_Universe}$), we have

$$Bel(\mathbf{a}_5 \cdot (\top, \texttt{v\_beautiful}, a_3) \mid \top) \subseteq Bel(\mathbf{a}_4 \cdot (\top, \texttt{v\_beautiful}, a_3) \mid \top),$$

Hence $S_4 \prec_3 S_5$, i.e., S_5 is strictly preferred to S_4 according to $\preceq_3$.

Thus we see that applying the preference criteria $\preceq_1$, $\preceq_2$ and $\preceq_3$ *in that order*, S_5 emerges overall as the best explanation among the explanations given in Section 3.1 for the Miss Thailand dialogue. In fact it is shown in [2, 3] that, given any dialogue d, such a best explanation always exists (assuming d is explainable at all):

Theorem 4.4 ([2, 3])**.** *Suppose there exists an explanation for d. Then there exists an explanation $[\mathbf{a}_R \mid \blacktriangle_R]$ such that, for* any *other explanation $[\mathbf{a} \mid \blacktriangle]$ for d, the following hold:*
(i) $[\mathbf{a} \mid \blacktriangle] \preceq_1 [\mathbf{a}_R \mid \blacktriangle_R]$.
(ii) *If* $[\mathbf{a}_R \mid \blacktriangle_R] \sim_1 [\mathbf{a} \mid \blacktriangle]$ *then* $[\mathbf{a} \mid \blacktriangle] \preceq_2 [\mathbf{a}_R \mid \blacktriangle_R]$.
(iii) *If* $[\mathbf{a}_R \mid \blacktriangle_R] \sim_1 [\mathbf{a} \mid \blacktriangle]$ *and* $[\mathbf{a}_R \mid \blacktriangle_R] \sim_2 [\mathbf{a} \mid \blacktriangle]$ *then* $[\mathbf{a} \mid \blacktriangle] \preceq_3 [\mathbf{a}_R \mid \blacktriangle_R]$.
An algorithm to construct such an explanation is given in [2, 3]. This constructed explanation is called the rational explanation *there.*

In [2, 3] we propose that the rational explanation of d should be considered as a best explanation for d. If we run the algorithm mentioned in Theorem 4.4 on the Miss Thailand dialogue d_{Miss} then it returns explanation S_5 as output. Thus S_5 is the rational explanation for d_{Miss}, and thus can be considered a best explanation for d_{Miss}. So the machine can use S_5 as the basis on which to make inferences and predictions about the user's beliefs. Recall the belief trace of S_5 through d_{Miss} is

$$(\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe},\ \texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe},\ \texttt{v\_beautiful} \,\&\, \texttt{Miss\_Universe}).$$

Thus, using the rational explanation for d_{Miss}, the machine infers that the user did *not* believe that Miss Thailand was Miss Universe at the start of the dialogue, only that she is Miss

Universe *if* she is very beautiful. After being informed by the machine that she is indeed very beautiful, the user believes what the machine says is true and then believes Miss Thailand is Miss Universe. If the machine were to then inform the user that Miss Thailand is not Miss Universe, the machine predicts (using the rational explanation) that the user will believe $\texttt{v\_beautiful} \,\&\, \neg\texttt{Miss\_Universe}$, i.e., the user will believe the new information but will still keep the belief that Miss Thailand is very beautiful. Finally note what happens if, instead of informing the user that $\neg\texttt{Miss\_Universe}$, the machine informs the user that Miss Thailand is *not* very beautiful. Then, again using $\mathbf{a}_5$ to denote the sequence $(\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe})$, we have

$$Bel([\mathbf{a}_5 \cdot (\top, \texttt{v\_beautiful}, \neg\texttt{v\_beautiful}) \mid \top]) = \\ (\texttt{v\_beautiful} \rightarrow \texttt{Miss\_Universe}) \,\&\, \neg\texttt{v\_beautiful}.$$

So in this case the user will be predicted to believe that Miss Thailand is not very beautiful (i.e., will believe the input given to him by the machine), but will no longer have any opinion on whether she is Miss Universe, since neither $\texttt{Miss\_Universe}$ nor $\neg\texttt{Miss\_Universe}$ are logical consequences of the above belief set.

5 Conclusion

We described how a machine, engaged in some dialogue with a user, might plausibly model that dialogue internally and make inference and predictions about the user's beliefs during the course of the dialogue by building some model of how the user forms and updates his beliefs.

There are several potential extensions of this work. One regards the assumption we made about the given dialogue d that between receiving the first input from the machine a_1 and the last a_n, the user did not receive *any* other information apart from the inputs a_i for $i = 2, 3, \ldots, n-1$. There are situations where it seems quite natural to *relax* this assumption. A good example is if the dialogue is conducted by the machine with the user over the course of more than one *session*. Between sessions, while the user is away from the machine, it is quite likely he will go about his life, absorbing extra information from *other* sources. Giving the machine the ability to guess what these "missing" inputs might be would give the machine an extra degree of freedom with which to explain d. Some initial results on this can be found in [5].

Also, recall that in this paper we concentrated entirely on phase 2 in our two-phase description of automated dialogue modelling from the start of Section 2. Obviously a satisfactory treatment of phase 1 is crucial, i.e., precisely *how* does the machine extract the sets B_i and D_i of propositional sentences from the user's natural language responses. Finally the focus of this work so far has been very much theoretical. However, as we said in Theorem 4.4, we *do* already have an algorithm which computes the rational explanation for an arbitrary dialogue d, although we have not yet implemented it. It would be nice to get an implementation of this algorithm up and running.

Acknowledgements

Thanks are due to Phattanaphong Chomnphuwiset for some helpful discussion which led to the Miss Thailand example.

References

[1] R. Booth, On the logic of iterated non-prioritised revision, in: *Conditionals, Information and Inference*, LNAI Vol. 3301, 86–107, 2005.

[2] R. Booth and A. Nittka, Reconstructing an agent's epistemic state from observations, in: *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*, 394–399, 2005.

[3] R. Booth and A. Nittka, Beyond the rational explanation, in: *Belief Change in Rational Agents*, Dagstuhl Seminar Proceedings, http://drops.dagstuhl.de/opus/volltexte/2005/332, 2005.

[4] H. Enderton, *A Mathematical Introduction to Logic*, Academic Press, Second edition, 2001.

[5] A. Nittka, Reasoning about an agent based on its revision history with missing inputs, to appear in: *Proceedings of the Tenth European Conference on Logics in Artificial Intelligence (JELIA'06)*, 2006.

[6] J. Paris, *The Uncertain Reasoner's Companion*, Cambridge University Press, 1994.