Mindreading Deception in Dialog

Alistair M.C. Isaac[a], Will Bridewell[b]

[a]*Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19104*
[b]*Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305*

Abstract

This paper considers the problem of detecting deceptive agents in a conversational context. We argue that distinguishing between types of deception is required to generate successful action. This consideration motivates a novel taxonomy of deceptive and ignorant mental states, emphasizing the importance of an ulterior motive when classifying deceptive agents. After illustrating this taxonomy with a sequence of examples, we introduce a Framework for Identifying Deceptive Entities (FIDE) and demonstrate that FIDE has the representational power to distinguish between the members of our taxonomy. We conclude with some conjectures about how FIDE could be used for inference.

Keywords: mental state ascription, ulterior motives, lying, false belief

1. Introduction

Every productive interaction between humans depends critically on the process of mindreading: the attribution of mental states to other agents. If another person can successfully infer your beliefs and desires, then that person can interact with you strategically, anticipating and preparing for your actions to compete or coordinate. For artificial agents to successfully participate in complex social interactions, they also will need strategies for mindreading.

Ideal mindreading requires multimodal input, including utterances, eye movements, body gestures, galvanic skin response, and other features. However, humans routinely interact successfully in limited cue conditions. For example, in an internet chat session with a stranger, we lack access to facial expressions, body language, and past experiences with that person. Instead, we must attribute beliefs and desires to the stranger based solely on the words typed and our general knowledge of such situations. A limited cue scenario such as this offers a delineated testing ground for mindreading frameworks for artificial agents.

A considerable amount of work exists that attempts to infer mental state from dialog content alone (e.g., Cohen and Perrault, 1979; Carberry and Lam-

Email addresses: aisaac@sas.upenn.edu (Alistair M.C. Isaac), will.bridewell@stanford.edu (Will Bridewell)

bert, 1999), but the bulk of this literature assumes agents are cooperative and speak veridically. In contrast, this paper introduces a general framework for mindreading from dialog with the express goal of detecting deception. More specifically, we take as our primary focus the problem of distinguishing between different types of deceptive and sincere speech. Our motivation is the observation that the correct response to an utterance differs depending on its categorization (as a lie, a sincere statement of false belief, a veridical statement uttered with deceptive intent, etc.). One might confront a liar, but gently correct a sincere but ignorant speaker. In some contexts, one might even decide to participate in a falsehood if it is strategically efficacious.

To crystallize the problem, we concentrate on the special case in which an agent believes some proposition P and his interlocutor utters not(P). In this situation, mindreading is critical for determining how beliefs about the other agent should be revised and deciding what further actions to take. Does the speaker utter not(P) sincerely? Is he lying? If the former, the agent might act to educate the speaker by providing his evidence for P; if the latter, we claim that the agent must infer the speaker's ulterior motive. The need to correctly calculate response on the basis of a conflict between a speaker's utterance and the hearer's belief appears in many conversational scenarios including police interrogations, court testimony, physician–patient interactions, political debates, and even water-cooler talk.

The literature on deception emphasizes a number of fine grained distinctions for characterizing different attitudes an agent might have toward an utterance. In addition to lying and distinct from the sincere categories of false belief and ignorance, there are other less well studied forms of deception. For example, *paltering* involves speaking truthfully with the intent to deceive (Schauer and Zeckhauser, 2009), as when a car dealer (veridically) emphasizes the quality of a car's wheels to distract the buyer from a problem with the engine. *Bullshitting* involves speaking without either knowing or caring about the truth value of one's utterance (Frankfurt, 2005). A less discussed, but also interesting example is *pandering*, when the speaker does not care about the truth value of their utterance, instead speaking it solely because they believe the listener desires to hear it. Whereas the philosophical literature on deception has focused primarily on the moral status of these classes of utterances, we focus on how to distinguish among them in a limited-cue, conversational situation.

Our strategy differs from previous work in that we (1) categorize a broad variety of deceptive states within a unified framework, (2) emphasize the importance of ulterior motives rather than "intent to deceive," and (3) propose a representation designed to be rich enough to support the detection of deception. The claim that deception and sincere discourse are qualitatively distinct is uncontroversial. We go a step further and assert that attempts to treat bullshitting, lying, pandering, and paltering as a single activity will lead to as much confusion as treating false belief, ignorance, and veridical speech as one and the same. The efforts of Sakama et al. (2010) to logically define and analyze lying, bullshit, and paltering—which they call "deception"—support our assertion. However, their logical analysis turns on acceptance of the condition that liars intend the hearer to believe their fallacious utterance, which has been hotly disputed (Mahon, 2008). We sidestep that debate by relying on the more general concept of an ulterior motive: intuitively, a goal with higher priority than those goals implied by the conversational context. The intent to deceive may follow from such a goal, but the ulterior motive itself ultimately determines action. These considerations motivate our Framework for Identifying Deceptive Entities (FIDE). We argue that FIDE includes features both necessary and sufficient for a mindreading system to detect deception.

We structure the rest of this paper around FIDE. In the next section, we develop a classification of deceptive states, illustrating our distinctions with multiple interpretations of a short dialog. We argue for the necessity of representing the goals and beliefs of other agents, including their potential ignorance and ulterior motives, so as to fully characterize these distinctions. Section 3 introduces FIDE and our basic formalism for representing mental states. Section 4 illustrates the framework's expressive power by applying it to examples from section 2. Success at representing our examples demonstrates the sufficiency of FIDE for capturing our distinctions. We then discuss strategies for inferring deceptive states within a system that implements this framework. Finally, we summarize our argument and anticipate future research.

2. An Example: The Water Cooler

To see more clearly the necessity of mindreading and reasoning about ulterior motives for dialog, consider the following simple exchange:

> Scene: *Bartleby & Bartleby, LLP*
> (Jones and Pratt stand next to a gurgling water cooler.)
>
> Jones: So, I hear Smith is going to be promoted to VP.
>
> Pratt: That's what you get for kissing old man Bartleby's ass.

Jones' response to Pratt will depend critically on his own mental attitude toward Smith's promotion. For the sake of argument, assume that Jones believes Smith's promotion to be merit-based. How then should he interpret and respond to Pratt's assertion that it was the result of cronyism?

The most socially generous interpretation of Pratt's statement is as a straightforward instance of *false belief*. For example, Pratt may have observed but misinterpreted conversations between Smith and Bartleby, forming the sincerely held but inadequately justified and ultimately incorrect belief that the promotion was an act of cronyism. In this circumstance, Pratt's utterance may have no motive behind it other than the straightforward Gricean mandate to speak the truth. If Jones infers that motive, he might respond by offering Pratt evidence that Smith's promotion was merit-based, working with him to realize their shared goal of reaching the truth of the situation.

At the opposite end of the spectrum is the interpretation of Pratt's utterance as a full-blown instance of *lying*. For example, Pratt may know full well that

Smith's promotion is merit-based. His assertion to the contrary must then be based on some *ulterior motive*. For instance, Pratt may have the goal of bringing Jones to believe that Smith's promotion was undeserved. This goal is ulterior in the sense that it contradicts the default Gricean assumption that the purpose of conversation is to veridically convey truth about the state of the world. As we shall see, the presence of an ulterior motive most thoroughly characterizes deceptive speech, not the speaker's attitude toward the truth value of his utterance. If Jones identifies Pratt's utterance as a lie, his primary goal then becomes the discovery of Pratt's ulterior motive: why does Pratt want to manipulate him?

False belief and lying offer the most straightforward interpretations of Pratt's utterance, but other subtle variations exist. For instance, suppose Pratt lacks a firm view of his claim's truth value. In that case, he speaks from a position of *ignorance*. Unless he simply misspoke, this interpretation requires some backstory to be plausible. Perhaps in suggesting cronyism, Pratt intends to "fish for the truth" from Jones. Shame at his ignorance of the full situation prompts Pratt to say something in an attempt to elicit Jones' beliefs, not sway them. In this case, Pratt's utterance lacks an ulterior motive as he indeed cares about the truth of the situation. If Jones infers Pratt's ignorance, much as in the case of false belief, the natural response is to provide evidence that the promotion was merit-based.

Contrast pure ignorance with the case in which Pratt neither knows nor cares about the truth value of his utterance. Frankfurt (2005) calls utterances made with this attitude *bullshit*.[1] If Pratt makes his utterance without any regard for its truth, then he must have some goal unrelated to truth in making it: an ulterior motive. As traditionally conceived, the motives behind bullshitting tend to be more innocuous than those behind lying. Pratt may simply desire to appear informed about office politics, or he may simply enjoy arguing at the water-cooler. Crucially, however, sincere but incorrect utterances (ignorance and false belief) and bullshit motivate different responses from Jones. One might seek to educate an ignorant party, but there is little value in attempting to educate a bullshitter. The bullshitter's goal is neither to learn the truth nor to hide it, nor even to preserve the consistency of his beliefs. Instead, he seeks only to preserve the consistency of his narrative. Individual statements may or may not be true, but neither case will slow or spur the flow of bullshit more than the other.

Pandering is a special case of bullshit found in the political realm. This category resembles bullshit because the speaker does not care about the truth value of his utterance (although he may possibly know it). However, a specific ulterior motive distinguishes pandering, namely the speaker aims to boost his estimation in the mind of the listener by uttering something the listener wants to hear. If Pratt himself does not care about the reason behind Smith's promotion,

[1] Other analyses of our common sense notion of bullshit are possible, however, for instance "Cohen bullshit" (Cohen, 2002).

Table 1: The possible mental state attributions by the hearer after the speaker has uttered not(P). We use not(P) to signify the negation of the proposition P and ig(P) to signify ignorance of P, to be discussed further below. Bh = hearer's belief. BhBs = hearer's belief about speaker's belief. BhBsBh = hearer's belief about speaker's belief about hearer's belief.

Bh	**BhBs**	**BhBsBh**	ulterior motive (UM)	No UM
P	P	P	lying	misspoke
P	P	ig(P)	lying	misspoke
P	P	not(P)	lying / pandering	misspoke
P	ig(P)	P	bullshit	ignorance
P	ig(P)	ig(P)	bullshit	ignorance
P	ig(P)	not(P)	bullshit / pandering	ignorance
P	not(P)	P	paltering	false belief
P	not(P)	ig(P)	paltering	false belief
P	not(P)	not(P)	paltering / pandering	false belief

but believes Jones to be bitter or jealous about it, then he may claim that the promotion was the result of cronyism with the goal of increasing his standing in the eyes of Jones.

This example highlights the expressive power our model of mental states will need to characterize the nuanced goal attributions required to mindread in a dialog context. From Smith's perspective, it is necessary to think not only that Pratt believes the promotion to be merited but also that Pratt believes that Jones believes (or wants to believe) that it resulted from favoritism. We will need to be able to handle arbitrary embedding of belief and desire operators to characterize cases such as this, where Pratt believes that Jones desires that Pratt believes that P.

A final form of deception worth considering is *paltering* (Schauer and Zeckhauser, 2009), which occurs when a speaker knowingly states the truth but with an ulterior motive. Typical cases involve the speaker uttering a minor truth to distract from the point of primary interest. For example, a used car dealer might mislead a buyer by pointing to a car's virtues to distract from its failings. To interpret the present example as an instance of paltering, we distinguish the pragmatic meaning of Pratt's utterance (i.e., Smith's promotion resulted from cronyism) from its colloquial meaning (i.e., Smith flattered old man Bartleby). With this in mind, suppose both that Pratt knows Smith's promotion is merit based and that Smith frequently flatters Bartleby. By accentuating Smith's flattery in the conversational context, Pratt attempts to mislead Jones into concluding that the promotion resulted from cronyism rather than merit.

Each of these interpretations of the Bartleby example has a different signature in Jones' attribution of mental states to Pratt. Table 1 summarizes the distinctions introduced here. To map this table onto the water cooler example, (a) Jones is the hearer, h, (b) Pratt is the speaker, s, and (c) the proposition "Smith's promotion was merit-based" is P. Note that the truth value of P is immaterial here; in analyzing how the hearer attributes mental states to the speaker, it is what the hearer believes that matters, not whether those beliefs are correct. The presence of an ulterior motive marks the distinction between a deceptive and a misguided utterance. The exact type of deception then depends upon the valence of belief in P attributed to the speaker by the hearer.

3. FIDE: A Framework for Mindreading

As the previous section demonstrates, mindreading is necessary for dialog in general and lie detection in particular. We propose a framework for mindreading that we call FIDE, Framework for Identifying Deceptive Entities. FIDE is limited to the synchronic representation of complex mental states and will need to be supplemented with an inference procedure to support dynamic reasoning about changes in mental state attribution. We discuss some heuristics for dynamic mindreading in section 5, but FIDE itself makes no assumptions about how inference is implemented. Summarizing the implications of the previous section, to effectively represent deceptive states, FIDE must be able to represent (1) beliefs about self and others, (2) the known ignorance of others, and (3) ulterior motives.

The central component of FIDE is its explicit attribution of mental states to other agents, preparing it for socially aware inference. The framework is organized around agent models and designates one of these as the model of self. All other agent models are nested under this model so that we might start from Jones's perspective and represent his model of Pratt or, more deeply, represent his model of Pratt's model of Bartleby. Each of these models contains mental content that we take to be propositional in character, but we also allow for properties useful during inference, such as justifications and entrenchment.

Each agent model contains two kinds of mental content: beliefs and goals. Beliefs are taken to be true by the agent, and assessable as true or false by inspecting reality. Goals represent desires about the state of the world and may be satisfied by complementary beliefs. Whether a proposition constitutes a belief or a goal is determined by the partition in which it is located. Importantly in the case of deception, an agent may have goals for another agent's beliefs. Furthermore, we distinguish a special class of goals, which we call "ethical goals," characterized by two features: (1) they are ascribed by default to all agents; (2) they are defeasible. An ethical goal is defeated when an agent prioritizes an ordinary goal over the ethical goal.

To illustrate, during Jones' conversation with Pratt, Jones maintains a model of his conversational partner. The belief partition of his self model links to a new model representing his beliefs about Pratt's mental state. If he attributes to Pratt beliefs or goals about the mental state of another agent (whether it be Jones or some third party), he creates a third agent model with links to the relevant partition in his Pratt model. The process can be iterated arbitrarily deeply as needed. Syntactically, goal and belief operators can embed to access the relevant content. For instance, we write (B Jones (B Pratt P)) to refer to Jones' belief that Pratt believes that P, and (B Jones (G Pratt (B Jones not(P)))) to indicate Jones' belief that Pratt has a goal that Jones believe the negation of P.

As a first approximation of mindreading, we take a cue from the ViewGen approach (Ballim and Wilks, 1991) which was adapted by Lee (1998) to distinguish lying from false belief. In particular, we assume that child agent models inherit the beliefs, but not the goals, of their parent agent through a process of default ascription. In the case just described, FIDE would model Pratt as implicitly believing everything that Jones does unless there is reason to think otherwise. Only those beliefs of Pratt's that differ from Jones' are explicitly stored in the Pratt model. Similarly, we assume that all agents share the same inference mechanisms and rules. This feature lets Jones simulate Pratt's reasoning using his own mechanisms and knowledge.

To represent scenarios in which Jones believes that Pratt's epistemic state differs from his own (or vice versa), we must be able to block the inheritance of beliefs from the parent model to the child model. In FIDE, inheritance is blocked when propositions contradicting those in the parent's belief partition are assigned to the child's belief partition. For example, if Jones believes that Pratt falsely believes not(P), then we write (B Jones P) and (B Jones (B Pratt not(P))), blocking inheritance of P into the Pratt model. To represent ignorance by this method, we introduce a special operator, ig(), which we treat in the same way as negation. If Jones believes Pratt is ignorant of P, then (B Jones P) and (B Jones (B Pratt ig(P))), and again inheritance of P into the Pratt model is blocked. Since beliefs are accessible by the agent that holds them, Pratt should be understood as aware of his own ignorance. In other words, FIDE represents known–knowns and known–unknowns, but not unknown–unknowns.

The final piece of FIDE is its ability to represent ulterior motives. Conceptually these are the goals of an agent that differ from its stated or implied goals. For instance, consider a patient in a hospital emergency room who requests Vicodin to alleviate a severe headache. The patient's stated goal is to get relief from a specific pain, but Vicodin contains an addictive narcotic agent. The patient may have an ulterior motive to procure Vicodin for recreational purposes or to distribute it illegally. In deciding whether the patient's request for Vicodin is deceptive, the physician need not determine the specific ulterior motive. Rather, the crucial task is to determine whether the patient is being direct or pursuing some unstated agenda.

Ulterior motives are intimately connected to behavioral norms, as disingenuous agents may use normative expectations to hide their actual goals. The notion of ethical goals lets FIDE support the ability to reason about behavioral norms. In the context of lie detection, we are primarily interested in the behavioral norm which Grice called the maxim of quality, i.e. the injunction to only utter statements which one believes to be true. In FIDE, we represent Grice's maxim with the distinguished ethical goal T, the goal to only utter the statement that-P if P can be found in the agent's belief partition. Since T is an ethical goal, any ordinary goal may defeat it, which we represent with the *defeat* relation that operates over mental content. Such a goal is "ulterior" in the technical sense that it lies beyond the pragmatic norms of conversation, but

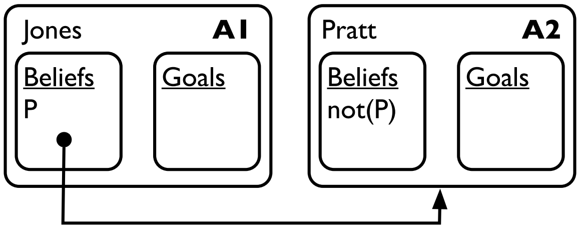

Figure 1: False belief. A1 and A2 are agent models with separate belief and goal partitions. Here, Jones believes that Pratt falsely believes not(P).

it may still play this role even if it is not "ulterior" in the colloquial sense of being hidden or secret (although typically it will be).

We represent that goal X defeats T for agent Agt as (G Agt defeat(X, T)). That is, the derived goal defeat(X, T) appears in an agent's goal partition whenever X is an ulterior motive for that agent. This goal licenses the agent to perform actions that violate T so long as they satisfy X. Importantly, defeat does not necessarily imply conflict. That is, some action that does not contradict T may also satisfy X. The patient requesting Vicodin for illicit purposes may nevertheless really be in severe pain. Therefore, we take the position that any X that defeats T is an example of an ulterior motive *regardless of whether the agent's actions violate T*.

In the previous section, we argued that these features—beliefs about other agents, ignorance, and ulterior motives—are necessary for distinguishing among several types of deception, and therefore for detecting deception in intelligent agents. In the following section, we demonstrate the sufficiency of FIDE for representing the mental states associated with each of the types of deception discussed in section 2, thus confirming its representational adequacy.

4. Representational Power of FIDE

Section 2 began with a short dialog that helped illustrate seven different types of deceptive or ignorant speech depending on context. Table 1 summarizes these different categories and provides recipes for representing them in FIDE. To illustrate, we will discuss three examples in detail, and leave the rest as an exercise for the reader.

We treat Jones as the "top agent," using FIDE to model his perspective of the scenario including his beliefs about the beliefs and goals of Pratt. For the remainder of this section, let P stand for the proposition that "Smith's promotion is merit-based." Since Jones believes P, we find it in his belief partition. When Jones meets Pratt for conversation, he adds a new partition to his workspace representing Pratt's mental state. If Pratt utters not(P) honestly, then by Jones' lights, Pratt falsely believes not(P). We represent this in FIDE by adding not(P) to the Pratt belief partition. Figure 1 illustrates this case with A1 being Jones' model of self and A2 being Jones' model of Pratt.

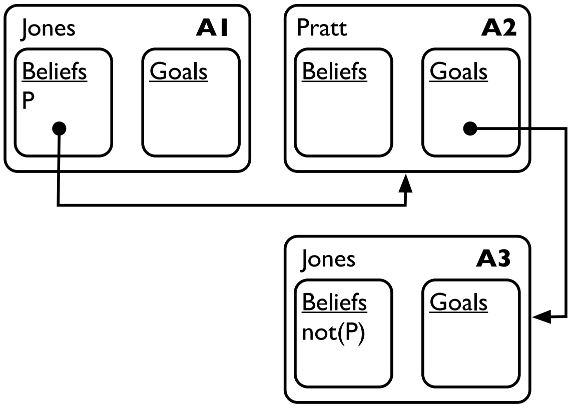

Figure 2: Lying. A1, A2, and A3 are agent models. Here, Jones believes that Pratt has lied, wanting him to believe not(P).

Importantly, this representation is purely synchronic. If Jones represents Pratt as in figure 1 *before* he hears Pratt utter not(P), then hearing Pratt's utterance does not change Jones' belief state. If Jones infers through some rule that Pratt has spoken honestly on the basis of false belief *after* hearing Pratt's utterance, then he will need to add not(P) to the belief partition of his Pratt model to block the default inheritance of P from his beliefs to Pratt's.

Now consider the simplest form of deceptive speech: lying. In this scenario, Pratt believes P but utters not(P) to deceive Jones. As with the previous example, we are interested in representing Jones' synchronic belief state. Figure 2 depicts Jones' beliefs, with A1 being Jones' model of self, A2 being his model of Pratt, and A3 being his model of Pratt's goals for Jones' mental state. Note that the Pratt model, A2, represents belief in P by default ascription. Although Pratt believes P, he has the goal of bringing Jones to believe not(P), a fact depicted by the link from A2's goal partition to A3 with not(P) in the belief partition. The link to agent model A3 represents a low level goal of agent A2, call it X. X satisfies the definition of an ulterior motive as it motivates utterances that contradict T, the maxim of quality (whether or not these utterances are actually made!). In this model then, (G A2 defeat(X, T)) can be derived.

Note that we are representing here the special case of lying in which the liar explicitly intends the listener to believe the fallacious utterance. This intent condition is controversial; for example, one might lie with the intent to deceive an eavesdropping third party, not one's conversational partner (see Mahon, 2008, for a survey of arguments against the intent condition). FIDE has the power to depict this special case,[2] but it can also depict more general instances of

[2] FIDE's ability to represent a wide variety of deceptive interactions lets us consider several such cases. For instance, as one reviewer suggested, a typical liar may believe the hearer to be ignorant of the truth value of his utterance. FIDE can easily represent this analysis from the speaker's perspective by stipulating (B Pratt (B Jones ig(P))) and from the hearer's perspective as the second row of table 1.

lying. For us, the defining features of a lie are (1) a contradiction between the speaker's belief and utterance and (2) presence of an ulterior motive (i.e., a goal that defeats T).

While the explicit intent to deceive constitutes an ulterior motive, much more abstract ulterior motives are also possible. Even if the intent to deceive is present, it will in general follow from some hierarchically more important goal. In the case of the Jones and Pratt dialog, Pratt may wish to incite a rivalry between Jones and Smith to open the door for his own promotion. In the Vicodin example mentioned in section 3, either the goal to get high or the goal to sell Vicodin on the black market might motivate the patient to lie. Both the more abstract motive and the more direct goal to generate a false belief in the hearer satisfy the definition of an ulterior motive. In the context of inference, awareness of the abstract goal on the part of the top agent might let him infer that the speaker is lying or, conversely, awareness that the speaker has an explicit intent to deceive might cause the top agent to infer a new goal, namely to discover the abstract motive behind this low level intention.

As a final example, consider pandering. In table 1, we illustrated an analysis of pandering in terms of tertiary structure: the top agent believes his interlocutor believes the top agent believes his utterance, and furthermore the utterance was made on the basis of this belief so as to ingratiate himself with the top agent. A more subtle analysis can be provided if we use a quaternary structure, easily accomplished with FIDE and shown in figure 3. Recall that the speaker who panders says what he believes the listener wants to hear, but "wants to hear" is plausibly interpreted as "desires the speaker to believe." Jones, in believing that Pratt panders to him by saying not(P), believes that Pratt believes that Jones desires Pratt to believe not(P). On this analysis, the actual belief P explicitly represented in A1 and implied by default ascription in A2 and A3 is irrelevant. Only the speaker's beliefs about the listener's goals are important.

Thus, whether the politician who says, "Michigan has beautiful trees," at a fundraising rally in Detroit is pandering depends not on the attendees' beliefs, not on his beliefs, and not even on his beliefs about the attendees's beliefs. Rather, the crucial question is, does he believe the attendees want him to believe the trees in Michigan are beautiful? If he makes the remark on the basis of that belief, then it constitutes pandering whether he or the attendees actually believe Michigan's trees are beautiful. This is a case where the goal to pander—to say what the hearer wants to hear—is indeed an ulterior motive, and defeats T. Nevertheless, since the speaker's own beliefs are irrelevant, T may be accidentally satisfied because the action which satisfies the ulterior motive may (incidentally) satisfy T as well.

Pandering and lying are relatively easy targets for detection since there is a particular signature to the pattern of mental state attributions which generates the lowest level ulterior motive. A much more challenging example is bullshit. An utterance is bullshit if the speaker neither knows nor cares about its truth value. Not caring about the truth value of P will defeat T, but what is the ulterior motive? Unlike the cases of lying and pandering, there is no distinctive low level goal here definable in terms of the pattern of mental state attribu-

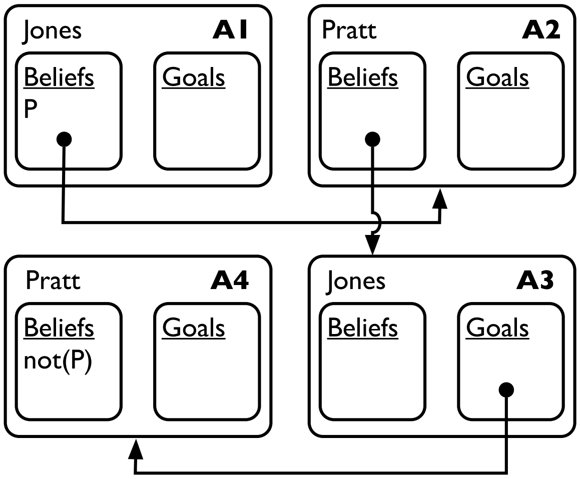

Figure 3: Pandering. A1, A2, A3, and A4 are agent models. Here, Jones believes that Pratt's declaration of not(P) is pandering because Pratt thinks Jones wants Pratt to believe not(P).

tions between hearer and speaker. In addition to the presence of ig(P) in his belief partition, there is not much more that can be inferred directly about the bullshitter other than simply that he possesses *some* goal that defeats T.

5. Strategies for Inference

We designed FIDE to represent mental state attributions synchronically. To dynamically reason about changes in mental state attribution, the framework requires an inference procedure. Our intention at this point is not to identify particular inference rules, which we take to be outside the scope of this paper, but to point out some characteristics those rules might evince and some strategies that one might follow when developing them. We begin by identifying the two principle questions whose answers are necessary for detecting deception, follow this with scenarios where some categories of deception are ruled out by context, and end with a recipe for inferring mental states.

The fundamental lesson of table 1 is that the question of mindreading when a speaker states some proposition that contradicts the hearer's beliefs breaks down into two distinct parts:

1. Does the speaker have an ulterior motive?
2. What is the valence of the speaker's attitude toward the proposition?

If the top agent has evidence for an answer to one of these questions, then he can focus on discovering evidence for the other.

For instance, if Jones already knows Pratt has a goal S to sow strife at the office and (B Jones (G Pratt defeat(S, T))), then he knows that Pratt is being deceptive and need only determine whether Pratt has evidence for cronyism. The nature of the evidence, if any, will establish the form of the deception.

Conversely, if Jones knows Pratt was absent the previous month and therefore has no inside knowledge about Smith's promotion, represented as (B Jones (B Pratt ig(P))), then he can turn his attention to the question of ulterior motive: is Pratt innocently fishing for more information or does he have some T-defeating goal, and is thus bullshitting?

This example brings us to another important point: we can dramatically simplify the problem by allowing plausible default assumptions. In many contexts, one can safely ignore the categories of misspeaking, ignorance, and paltering (schematically, imagine crossing these categories off in table 1). Thus from the valence of the belief, one can directly infer the presence or absence of an ulterior motive. From the assumption that there is no ulterior motive, one derives false belief. If there is evidence for an ulterior motive, then the hearer can turn his attention to determining whether the speaker believes P or ig(P). In some contexts, this further problem can be simplified by taking lying as the default. In water-cooler conversations, bullshit may be as common as lies; on the witness stand, however, bullshit seems an unlikely analysis.

Another helpful heuristic dictates that the top agent infer only as much detail about the speaker's deceptive behavior as is necessary to guide action. Different types of ignorant and deceptive speech require different actions in some contexts, but not in others. If Jones and Pratt are peers, Jones may simply choose to ignore Pratt's comments once he determines there is an ulterior motive—exactly how or why Pratt is deceiving him may be irrelevant. Suppose, however, Pratt is Jones' superior. Then determining Pratt's exact form of deceptive speech may be important. If Pratt is merely bullshitting, Jones may ignore the comment, but if Pratt is lying, then Jones may take his comment as strategically significant: does this mean Pratt hopes for more brown-nose behavior from Jones? That he expects Jones to stay silent when the matter is brought up at an upcoming staff meeting? Further inference is required.

In some cases, determining the type of deceptive speech will not be enough. Consider again the Vicodin example. In the emergency room, the key issue is whether the patient is lying: if not, then he should be treated; if so, he should be turned away. Suppose, however, the exchange occurs with a general practitioner. If the general practitioner determines the patient is lying, he may still need to delve deeper and infer the full ulterior motive. A patient trying to procure Vicodin for the black market may be sent away, but if the motive is addiction, then the physician should recommend a treatment program.

A final implication of the discussion in previous sections is that special rules for deceptive mental states with distinctive signatures may aid inference. In particular, recall pandering. As discussed above and shown in figure 3, pandering involves a distinct signature with quaternary embedding of the form belief – belief – goal – belief, with the uttered proposition found in the inmost level of nesting, A4's beliefs. An inference system might take advantage of this analysis by including a rule to infer the existence of an ulterior motive for the agent represented by A2 whenever this distinctive signature is encountered.

Combining these insights produces a rough recipe for inferring deceptive mental states using FIDE:

1. Omit all categories of ignorance and deception permitted by the conversational context.
2. Determine the specificity of mental state attribution required to generate one's next action.
3. Does the current signature fit any special inference rules?
4. Is there an ulterior motive?
5. What is the attitude of A2 toward the proposition P?
6. If mental state attribution is still not specific enough to generate an action, posit either an ulterior motive or a propositional attitude (whichever is missing) and return to 3.

Although this recipe is far from a well defined computational procedure or full set of inference rules, we believe that it provides a reasonable starting point for efforts in those directions.

6. Conclusion

This paper introduced FIDE, a framework for mental state attribution in the context of dialog systems. After discussing some examples of the type of mindreading required to effectively engage in realistic dialog, we introduced a taxonomy of ignorant and deceptive states. We found that deception correlates with the presence of an ulterior motive, but the exact type of deceptive state depends upon both the speaker's own attitude toward the spoken proposition and his attribution of beliefs and goals about that proposition to the hearer. We demonstrated the adequacy of FIDE for capturing the distinctions in our taxonomy, then concluded with a brief discussion of some of the insights for inferring mental states suggested by our analysis. Although the particulars will depend upon the inference procedure used, we were able to suggest a basic recipe for inferring deceptive states using FIDE in a dialog system.

In the future, we hope to further explore the modeling power of FIDE, especially with respect to the use of ethical goals. For instance, some taxonomies of deceptive speech consider violations of Gricean maxims other than the maxim of quality (e.g., Gupta et al., 2012). How much representational power would we gain by explicitly including additional conversational maxims on the list of ethical goals? We also intend to consider other general default ethical assumptions (e.g., maxims not to steal or not to murder). Would explicitly including these assumptions enable more powerful mindreading in a dialog context? We are also intrigued by some examples of complex deceptive speech which we cannot yet fully characterize. For instance, a caretaker suffering from Munchausen by proxy attempts to deceive physicians into believing her ward suffers from some chronic disease (Awadallah et al., 2005). This case is interesting from a modeling standpoint as there appears to be an ulterior motive, but there is no theory about what that motive itself is. Crucially, sufferers from Munchausen by proxy themselves lack access to the motive, creating grave difficulties in detecting and reasoning about the condition (Squires and R. H. Squires, 2010).

Mindreading deceptive agents is a challenging task, but one required of any realistic dialog system. We hope to have made some progress on simplifying this problem with the analysis presented here.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1028130 and by the Office of Naval Research under Contract No. ONR-N00014-09-1-1029. We thank Paul Bello, Amar K. Das, Wei-Nchih Lee, and two anonymous reviewers for their helpful comments.

References

Awadallah, N., Vaughan, A., Franco, K., Munir, F., Sharaby, N., Goldfarb, J., 2005. Munchausen by proxy: a case, chart series, and literature review of older victims. Child Abuse & Neglect, 931–941.

Ballim, A., Wilks, Y., 1991. Artificial believers. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Carberry, S., Lambert, L., 1999. A process model for recognizing communicative acts and modeling negotiation subdialogues. Computational Linguistics, 1–53.

Cohen, G. A., 2002. Deeper into bullshit. In: Buss, S., Overton, L. (Eds.), Contours of Agency: Essays on Themes from Harry Frankfurt. MIT Press, Cambridge, MA, pp. 321–339.

Cohen, P. R., Perrault, C. R., 1979. Elements of a plan based theory of speech acts. Cognitive Science, 177–212.

Frankfurt, H. G., 2005. On bullshit. Princeton University Press, Princeton, NJ.

Gupta, S., Sakamoto, K., Ortony, A., 2012. Telling it like it isn't: a comprehensive approach to analyzing verbal deception. In: Paglieri, F., Tummolini, L., Falcone, R., Miceli, M. (Eds.), The Goals of Cognition: Festschrift for Cristiano Castelfranchi. College Publications, pp. 567–600, also available from `http://cogsys.ihpc.a-star.edu.sg/publications`.

Lee, M., 1998. Belief, rationality, and inference: a general theory of computational pragmatics. Doctoral dissertation, Department of Computer Science, The University of Sheffield, Sheffield, England.

Mahon, J. E., 2008. The definition of lying and deception. In: Zalta, E. N. (Ed.), The Stanford Encyclopedia of Philosophy, fall 2008 Edition.

Sakama, C., Caminada, M., Herzig, A., 2010. A logical account of lying. In: Proceedings of the Twelfth European Conference on Logics in Artificial Intelligence. Helsinki, Finland, pp. 286–299.

Schauer, F., Zeckhauser, R., 2009. Paltering. In: Harrington, B. (Ed.), Deception: From Ancient Empires to Internet Dating. Stanford University Press, Stanford, CA, pp. 38–54.

Squires, J. E., R. H. Squires, J., 2010. Munchausen syndrome by proxy: ongoing clinical challenges. Journal of Pediatric Gastroenterology and Nutrition, 248–253.