Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium (FS-11-01)

# Recognizing Deception: A Model of Dynamic Belief Attribution

**Will Bridewell**
Stanford Center for Biomedical Informatics Research
Stanford University, Stanford, CA 94305

**Alistair Isaac**
Department of Philosophy
University of Michigan, Ann Arbor, MI 48109

**Abstract**

Social cognition is a key feature of human-level intelligence. However, social reasoning faculties are rarely included in cognitive systems. To encourage research in this direction, we introduce a practical, computational framework that enables *socially aware inference*. We demonstrate the framework's ability to model a common, complex, and under-investigated aspect of human social behavior: deception. Moreover, we show how a system implementing this framework could dynamically respond once it has detected a lie. We then discuss some of the challenges associated with deception, ending with an outline of future research directions.

## 1 Introducing. . . Deception

A bitter reality faces the intelligent system released into the world: *agents lie*. Incapable of judging a statement's veracity, the naif program believes everything it hears. The machine's guiding principle is that every interaction provides a clear and accurate signal of the state of the world. This assumption, perhaps unfortunately, lacks credibility. People routinely mask their beliefs and goals to impress others, to avoid humiliation, to manipulate their circumstances, and generally to control social situations. Intelligent systems incapable of detecting deception may respond inappropriately to common situations and leave themselves open to potentially disastrous consequences.

Consider a realistic situation initiated when a mother, K, brings her daughter, S, to the doctor's office. K explains that S has a history of chronic sinus problems and gastrointestinal pain. This is S's fifth visit to the doctor this year, and she has seen three physicians before this one. Moreover, S has undergone multiple tests and surgical procedures in the past four years, but K insists that her illnesses continue to reoccur. The doctor carries out a physical examination and finds no apparent medical problems, but K maintains that S routinely complains of headaches, sinus drainage, abdominal pain, and nausea. K pleads with the doctor to order new tests to help diagnose her daughter's condition. In reality, S is quite healthy, and K has fabricated her daughter's illnesses for years.

This scenario illustrates a particularly heinous form of deception directed at medical practitioners. The underlying condition is called "fabricated or induced illness" or "Münchausen by proxy" and is characterized by a caretaker who presents their ward for examination, insisting on nonexistent or induced symptoms and requesting unnecessary and potentially invasive medical procedures. The medical literature contains multiple attempts at explaining the behavior (e.g., seeking attention, asserting power or control), but acknowledges that detection alone can be difficult (Squires & Squires 2010). The physician is biased against detection both by professional ethics that encourage the most suitable treatment possible and by experience that suggests that parents typically do not manufacture illnesses for their children. Furthermore, the consequences of false accusations are high. These and other factors can lead medical professionals to overlook substantial evidence for fabricated symptoms.

Notably, physicians are counseled that in "making the diagnosis of a child as the victim of Münchausen by proxy, the motivation of the perpetrator needs to be assessed and understood" (Awadallah *et al.* 2005). Are ulterior motives driving the caretaker's behavior? If so, how should knowledge (or suspicion) of the motives alter the physician's mental state? In Münchausen by proxy, the motive may be the pursuit of a distinct psychological reward at the expense of one's child. But, unlike other cases where the motives are more apparent (e.g., to obtain prescription drugs, to avoid incarceration, to outwit one's peers) the actual motive of the caretaker is often a mystery. Even so, knowing only that an ulterior motive exists (i.e., a known unknown) may be sufficient for an agent to respond intelligently to the situation.

The need to reason about the motivations of other agents suggests that a deception savvy system must be able to represent not only its own beliefs, goals, and intentions, but also those of the agents with which it interacts. Returning to the example, the physician may have a belief that S is well and a belief that K believes that S is ill. However, deception necessitates the addition of a third layer to the model. Instead of holding the second belief, the physician may believe that K has a goal for the physician to believe that S is ill. That is, K would have the intention to lie in order to produce a false belief in the doctor who will then satisfy her hidden goal. This description follows the definition of lying provided by Schauer and Zeckhauser (2009) that includes (1) the intent to deceive by a speaker, (2) the expression of a false statement by that speaker, and (3) the production of a misrepresentation of reality in a listener.

The social act of deception offers a critical challenge that any approach to mental state ascription (mind-reading) aspiring to model the richness of human interaction must address. In this paper, we present a framework for mind-reading and claim that it can (a) represent dialogs with both honest and deceptive agents and (b) dynamically adjust when lying is detected. The next section describes this framework and motivates its advantages in the context of related approaches. We illustrate the richness of the representation by showing that it can characterize naive and skeptical agents. We then substantiate the framework's flexibility using examples wherein an agent's attitude dynamically shifts from trust to suspicion. Finally, we discuss the limits of this research and future directions.

## 2 Agents and Their Simulacra

Mind-reading is the process of attributing mental states to other agents. For a physician to detect Münchausen by proxy, he must perform complex inferences about the perpetrator's mental states. This process requires reasoning about the caretaker's beliefs, her goals, and her intentions. The physician's model must be dynamic, that is, have the ability to change its attribution of mental states to the perpetrator as evidence accumulates. Furthermore, inferences over the model must take place in real time if they are to guide dialog performance.

### A Few Standard Frameworks

Scientists and philosophers usally develop models of mind-reading against the general backdrop of theory of mind, which analyzes an agent's mental state in terms of beliefs, desires, and intentions. Researchers across fields have explored theory of mind, with the most extensive work appearing in the psychological literature. There, behavior on tasks that involve increasingly sophisticated reasoning about other agents marks the stages of child development (e.g., Wellman 1990). In artificial intelligence, theory of mind has inspired architectures for reasoning about other agents (Rao & Georgeff 1995). However, these approaches provide a static framework for describing other agents and must be extended with dynamics to develop a real-time model of reasoning about other agents.

One approach to the dynamics of belief can be found in the literature on belief revision (Alchourrón *et al.* 1985). This work explores procedures for updating databases or, more generally, theories as new evidence accumulates. Although this approach is suitably dynamic, it does not explicitly model the mental states of other agents. Information sources can be treated with varying degrees of skepticism or trust, but this assessment is not inferred dynamically from the incoming signals themselves as it is in the Münchausen by proxy example.

A related approach which does explicitly model the belief states of multiple agents can be found in the literature on dynamic epistemic logic (van Ditmarsch *et al.* 2008). This approach can explicitly represent the modalities of belief and knowledge for multiple agents in its syntax. However, much of the work in this tradition, including the fragment called public announcement logic (Baltag *et al.* 1998), analyzes changes in belief attribution on the assumption that utterances are veridical. Although recent work has explored variants of dynamic epistemic logic strong enough to define deception, there is as yet no approach to *detecting* deception within this tradition (e.g., Sakama *et al.* 2010).

Both strategies employ a Lewis-style semantics based on a structured set of worlds (Lewis 1973). Changes in the model involve reordering worlds, subtracting worlds, or changing the accessibility relation between worlds. The guiding principle behind this model is that new information reduces the number of possible worlds. Conceptually, this approach is a powerful analysis of the nature of information, but it is impractical as a concrete suggestion for a computationally tractable representational framework.

Our goal is to develop a framework which both (a) explicitly performs the task of mind-reading and (b) is computationally tractable. This means that the mental states of other agents will need to be explicitly represented so that they are available to inference mechanisms. It also means, however, that we need to restrict attention to inference procedures that can execute in real time. Our general strategy here is to take an approach opposite to that of possible worlds semantics. Rather than beginning with a representation of all possibilities that a system would then contract, we will only explicitly represent beliefs ascribed to other agents as they are generated during dialogue.

### A Socially Aware Framework

Our framework for mind-reading follows in the tradition of ViewGen, developed by Ballim and Wilks (1991). Their approach organizes propositional content into environments that represent either *topics* that beliefs may be about or *views* that an agent may have. Drawing from one of their examples (p. 155), ViewGen may have beliefs about John that he is male and six feet tall. Additionally, the system may believe about John that John believes that his cat is nice. That is, the topic environment of John has a nested view environment that represents John's perspective.

Importantly, nested views inherit the beliefs of their surrounding environments, which means that in this example, ViewGen also believes that John believes himself to be male and six feet tall. In principle, beliefs are stored in the outermost view to which they are ascribed and explicit negation in nested environments can block inheritance. Most importantly, the system incorporates a mechanism that dynamically adds beliefs to the correct environment and one that resolves questions of ascription by following well defined chains of inheritance.

Our framework differs from ViewGen in three key ways. First, our approach lacks support for topics, but we plan to incorporate representationally richer versions of these, similar to Cyc microtheories (Lenat 1998), in the near future. Second, ViewGen supports two forms of reasoning: percolation, which moves propositions from nested views outwards, and ascription, which controls the visibility of propositions within nested views. Both mechanisms are important, and here we assume ascription through default reasoning as in ViewGen. However, we supplement ViewGen's two

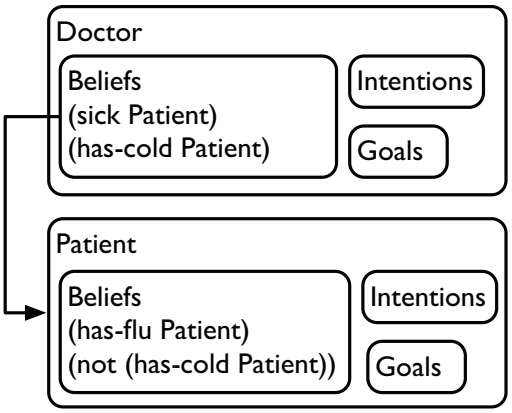


Figure 1: A doctor modeling the mental state of a patient. The outer boxes represent agent partitions, and the inner boxes represent the memory partitions that comprise each agent's mental state.

forms of reasoning with an abductive reasoner (Bridewell & Langley 2011) to create a general approach to *socially aware inference*. Third, we follow Lee's (1998) example and move beyond ViewGen's principal emphasis on belief to treat agents as complex structures that contain separate memory partitions for beliefs, goals, and intentions.

With this descriptive context in mind, we consolidate our framework as follows. An agent is represented by a mental space that is partitioned into models of agents, one of which is a privileged model of the self. Each agent model is partitioned into three sets of mental objects: beliefs, goals, and intentions. Beliefs are factual content that can be true or false. Goals are states of the world that the agent would like to bring about. Intentions are actions that may be realized in pursuit of a goal. Each primitive mental object consists of a literal that represents content and additional tags that indicate its source, its entrenchment, and other information. The modality of the tagged object is determined by where it is stored, and implicitly, intentions and goals are reflectively indexible as beliefs about intentions or goals.

When agent $A_1$ talks with another agent $A_2$, $A_1$ generates a model of its conversational partner and connects that model to $A_1$'s belief partition. This model represents $A_1$'s beliefs about $A_2$. $A_1$ can access its beliefs about $A_2$ through special operators such as (B $A_2$ <content>) or (B $A_2$ (G <content>)) where B, G, and I represent operators that access the belief, goal, and intention partitions of the specified agent. Since we index goals and intentions as beliefs, only the belief operator requires an agent argument.

As we mentioned, beliefs are inherited in nested agents by default. Therefore, when $A_1$ generates its model of $A_2$, only those beliefs that differ from those of $A_1$ are explicitly added to $A_2$. Additionally, all agents share the same inference rules, which are stored with the principal agent. This lets agent $A_1$ take $A_2$'s perspective by applying the same inference rules inside $A_2$'s model.

To illustrate this representation, consider the model in Figure 1. Here we have a principal agent, Doctor, who is reasoning about another agent, Patient. For simplicity, we omit goals and intentions, but show placeholders for their partitions. In this case, the doctor would ascribe (sick Patient) to both the doctor and the patient, (has-flu Patient) and (not (has-cold Patient)) only to the patient, and (has-cold Patient) only to the doctor. Note that the framework uses a default rule to ascribe (sick Patient) to the patient, but (not (has-cold Patient)) blocks the doctor's ascription of (has-cold Patient) to the patient. In general, we could also define constraints that assert mutually exclusive relationships to augment explicit negation and block ascription through default reasoning. Moreover, this example is limited to a case where the doctor has beliefs about the patient's mental state. As later examples will illustrate, an agent may also have goals for another's beliefs, goals, and intentions.

Although we will not employ this feature in our analysis of deception, notice that this approach supports basic self awareness. That means that the model also represents (B Doctor (sick Patient))), (B Patient (B Patient (has-flu Patient)))), and so forth. Since the patient lacks its own explicit model of the doctor, we assume that the patient ascribes all its beliefs to the doctor. More complex relationships are left as future research.

## 3 Modeling Deceptive Interactions

To demonstrate the representational power of this framework, we apply it to four medical situations of increasing complexity. Each hypothetical situation involves a physician and a patient in dialog during an emergency department visit. The physician is the principal agent. The first two situations are identical in content, but the physician prejudges the trustworthiness of the patient differently. Those two models exemplify extreme endpoints for a principal agent's attitude: complete naivete or complete skepticism. The third situation presents the same scenario, but here the physician starts out as a naive agent, detects that the patient is lying, and dynamically adjusts her mental state to support skepticism. The fourth situation revisits the Münchausen by proxy scenario described in Section 1.

### The Naif and the Skeptic

Consider the following dialog between a physician and a patient, which reflects an experience commonly encountered in emergency departments.

**Scenario 1**
*Physician*: "Tell me what's wrong."
*Patient*: "I have a terrible headache!"
*Physician*: "Have you tried pain relievers like Tylenol?"
*Patient*: "Oh no, doctor, I'm allergic to Tylenol."
*Patient*: "Last time, Vicodin really helped."
*Physician*: "..."

Superficially this interaction seems reasonable. The patient describes a symptom, warns the doctor of a relevant allergy, and mentions past successful treatment. However, two important pieces of domain knowledge reveal an interesting subtext. First, Vicodin is a trade name for a combination of hydrocodone, a controlled narcotic, and acetaminophen, a

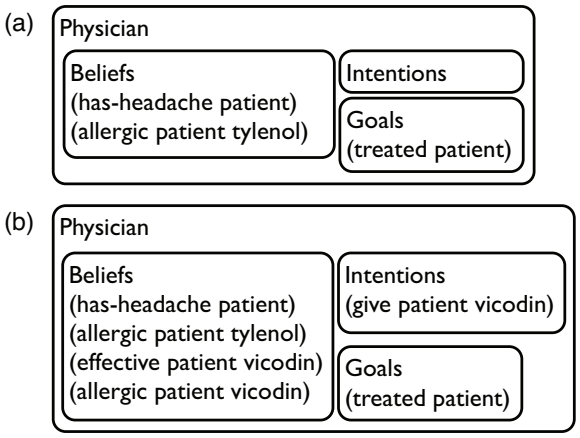


Figure 2: A snapshot of a naive agent's mental state in Scenario 1 (a) after the patient asserts an allergy to Tylenol and (b) after the patient requests Vicodin.

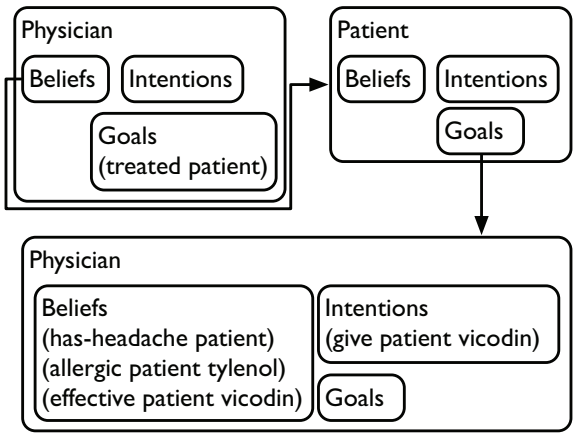


Figure 3: A snapshot of a skeptical agent's mental state after Scenario 1. Arrows connect agent partitions letting the physician explicitly store beliefs about the patient's mental state and the patient's goals for the physician's mental state.

common analgesic. Second, Tylenol is a trade name for acetaminophen. In this context, the patient's dialog suggests that he may be lying and engaging in drug seeking behavior to serve his addiction.

In Situation 1, we analyze this scenario by modeling the physician as a naive agent, one who makes a blanket assumption of veridicality. Figure 2 shows such an agent's potential mental state both after the patient states his allergy and after he requests Vicodin and the physician identifies a conflict. Here we are not modeling knowledge necessary to control the dialog, so the assumption of truth obviates the need for an explicit model of the patient. Instead, the physician collects facts about the world and processes them as she deems fit.

In Figure 2a, we see that the physician has accepted the patient's report of having a headache and established a goal to treat the patient. Additionally, the physician has accepted the patient's allergy. Figure 2b shows a later state, which now contains conflicting beliefs. By this point, the physician has accepted the patient's claim that Vicodin was an effective treatment and has formed the intention to administer Vicodin, (give patient vicodin). However, continued inference has produced the belief (allergic patient vicodin), which contradicts (effective patient vicodin). At this point, the physician must resolve the conflict in order to act appropriately. One potential solution is to err on the side of caution by removing the intention to administer vicodin and by suggesting another pain medication such as ibuprofen. Regardless of the response, this model cannot support a general approach to detecting and responding to the deception.

In Situation 2, we model the physician in this scenario as an intrinsically skeptical agent who never accepts the patient's statements at face value. For instance, such a situation might occur if, before seeing the patient, the physician met with a nurse who mentioned that the patient is a known addict who routinely visits the hospital seeking narcotics. In this case, the physician would not directly accept the patient's statements as beliefs. Figure 3 shows the potential mental state of this skeptical physician after the dialog. To emphasize the relationship between agents, we leave out some of the inference details.

In this model, the physician retains the goal to treat the patient, but skepticism has led to the creation of two new agent partitions. The first stores the physician's beliefs about the patient's mental state. This degree of removal is helpful when an agent holds false beliefs or when two agents disagree. As an example, the physician may have a record of the patient receiving acetaminophen without complications and might infer that the patient is confused. The second stores the patient's goals for the physician's beliefs. For instance, when the patient says, "I have a terrible headache!" instead of accepting this statement at face value, the skeptical physician explicitly interprets this as a belief that the patient would like the physician to hold: (B Patient (G (B Physician (has-headache Patient)))). This extra layer of representation lets the principal agent decide whether to trust each statement's content.

Figure 3 also demonstrates an advantage of our framework that plays a role in detecting deception. The intention (give patient vicodin) results from mental simulation, which involves an agent reasoning from an alternative perspective.[1] In this case, the skeptical physician assumed the perspective of the nested physician model, which also inherits the (treated patient) goal, and inferred the intention to administer Vicodin. Continued inference in this context would produce the mental state from Figure 2b. From that point, the skeptical agent must decide whether to accept the beliefs about allergies, the belief about effectiveness, or none of those and then act accordingly.

## Detecting Deception

So far, we have used our framework to model unrealistic agents. Neither the naif nor the skeptic can respond effec-

[1]This operation is directly related to the application of microtheories in Cyc (Lenat 1998).

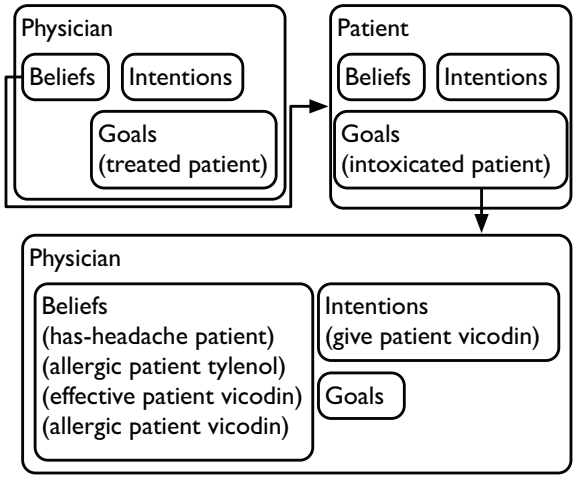


Figure 4: A snapshot of the physician's mental state after detecting deception.

tively in all social contexts. Regardless, most if not all frameworks for intelligent systems fall into the naive category. Humans fall somewhere between these two extremes, taking most people at face value and reacting appropriately when they realize that they are being deceived. As a result, we need to demonstrate that our framework can adapt its cognitive structures to respond to an uncovered or suspected lie.

Situation 3 revisits Scenario 1 by initially modeling the physician as a naive agent as in Situation 1. Eventually, this leads to the state shown in Figure 2b where the physician intends to give the patient Vicodin and has just inferred the contradictory belief that the patient is allergic to the drug. At this point, a system implementing the framework would enter a belief revision routine to update its mental state accordingly. This step may take a number of forms including removing one of the contradictory beliefs, altering the existing intention, or treating the interaction as deceptive. Here we explore this last case, shifting from our specific model of a physician to mechanisms in a cognitive system.

Before modeling the effects of the transition, we need to specify the conditions under which the system might detect deception. In this paper, we assume that a speaker lies to achieve one or more ulterior motives and that an agent may possess an ulterior motive if one can explain its statements with multiple, unrelated goals. In the current example, Vicodin and Tylenol satisfy many of the same goals due to their shared ingredient even though the former may be more potent than the latter. Over and above their shared effects, Vicodin as a narcotic achieves the unique goal of intoxication, which provides an ulterior motive for the patient's request.

Importantly, the mere potential for an ulterior motive is insufficient for one to detect deception without appearing paranoid or accusatory. However, once a system detects a contradiction in an agent's beliefs, the presence of an ulterior motive offers a heuristic for selecting a particular belief revision strategy. In this situation, the system responds plausibly by becoming skeptical toward the speaker. Recall that the skeptical agent protected its own beliefs by creating a separate partition to house the suspected lies. Here we demonstrate the shift from the memory partitions shown for the naive agent in Figure 2b to ones similar to those of the skeptic shown in Figure 3.

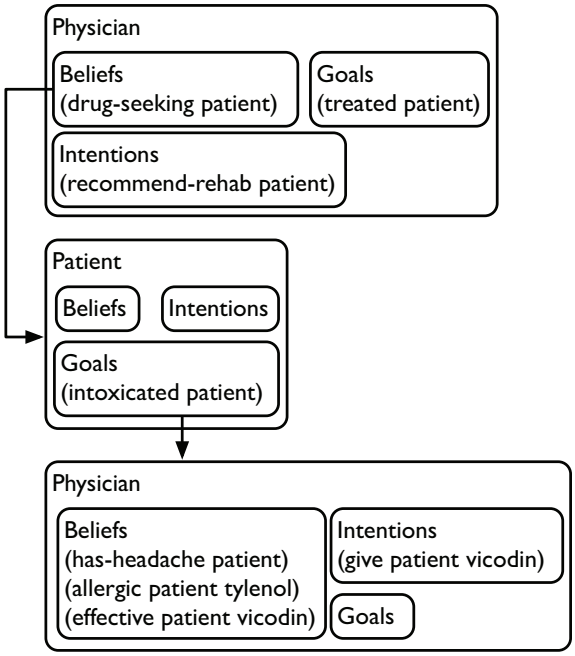


Figure 5: A snapshot of the physician's mental state after responding to deception.

The initial step in this procedure involves the hypothetical system transferring the mental objects supported by suspected lies from the principal agent's partition to the nested partition. This shift enables the system to reason about deception. After the transition process, the system is free to infer the patient's ulterior motive, (B Patient (G (intoxicated patient))). Figure 4 shows the results. (G (treated patient)) was left alone because it was supported by the context of the physician–patient relationship and not by the dialog content.

Once the structures are in place, the system can remove inferred beliefs that contradict stated ones and continue reasoning about the situation. Ultimately, the physician might reach the state shown in Figure 5, identifying the patient as a drug seeker and forming the intention to recommend a rehab clinic as a way to satisfy (G (treated patient)).

Having described how our framework can represent deception and be used by a system to detect deception, we now revisit our original motivating example in Situation 4. Detecting deception in Münchausen by proxy differs qualitatively from the drug seeking scenario because the caretaker's ulterior motives are unknown. It seems unreasonable to limit the detection of deception to cases with a specific ulterior motive. Consider the following dialog, which is an adaptation of the example from Section 1.

**Scenario 2**

*Physician*: "Tell me what's wrong."
*Caretaker*: "My daughter has chronic sinus problems."
*Caretaker*: "She always gets these headaches."

*Caretaker*: "She also has constant sinus drainage."
*Physician*: "What have you tried?"
*Caretaker*: "She's seen three doctors so far."
*Caretaker*: "She's had sinus surgery twice."
*(Physician carries out exam)*
*Physician*: "Your daughter appears fine right now."
*Caretaker*: "She *always* has headaches and drainage."
*Caretaker*: "Can you get an x-ray to see for certain?"
*Physician*: "..."

Figure 6 shows the physician's mental state at the end of the dialog. During the examination, the physician noticed that the patient seemed healthy, with no evidence of a headache or sinus drainage. Since the physical exam contradicted the caretaker's claims, the physician's mental model needed adjustment. In this case the physician retains beliefs supported by the exam, but moves the mental objects supported by the caretaker's statements into the secondary model. Even though there were no ulterior motives to influence belief revision, the model takes a form similar to one generated after detecting deception. This restructuring is plausible because it lets one represent conflicting information from two separate, but supposedly reliable, sources. The caretaker's empty—apart from the linked agent—goal partition indicates the absence of an ulterior motive. In this case, the caretaker inherits the goal (healthy patient) from the physician.

Furthermore, the physician's goal (healthy patient) is satisfied by an existing belief. This state justifies moving the intention (order-test x-ray patient) to the nested agent because taking that action, which suggests an unhealthy patient, is inconsistent with having a patient who is well. In addition, the caretaker's insistance that her daughter is unwell after the physician evaluated her health has led to the inference of (B Caretaker (I (insist-unhealthy patient)). Again, this alone is insufficient to infer deception even though there is continued disagreement between the agents. Instead, from the ascriptions (healthy patient), (B Caretaker (I (insist-unhealthy patient))), and (B Caretaker (G (B Physician (I (order-test x-ray patient))))), the physician could infer (B Caretaker (G (order-unnecessary-test x-ray patient))) without assessing the caretaker's honesty.

Finally, one may infer from a caretaker's goal to order unnecessary tests on their ward that the caretaker has Münchausen by proxy. That diagnosis then lends itself to detecting, post hoc, that deception has occurred—an inversion of the previous scenario, where the detection of deception assisted in the recognition of drug seeking behavior. Nevertheless, the caretaker's ulterior motive remains a mystery as it typically is neither to harm her ward nor to instigate tests for their own sake. Instead, the cause is likely related to a psychological need.

This section applied our proposed framework for socially aware inference to two scenarios involving lies. In each case, we used the framework to represent various mental states of a principal agent as it interpreted the content of a dialog. Furthermore, we identified boundary cases of skepticism and naivete and introduced a means by which a belief revision routine might detect and respond to a deceptive situation. The Münchausen by proxy scenario exemplifies an interesting challenge for systems that detect lies, absence of an explicit ulterior motive, and suggests that there are other challenges for researchers to explore in the future. The next section discusses some of these in limited detail.

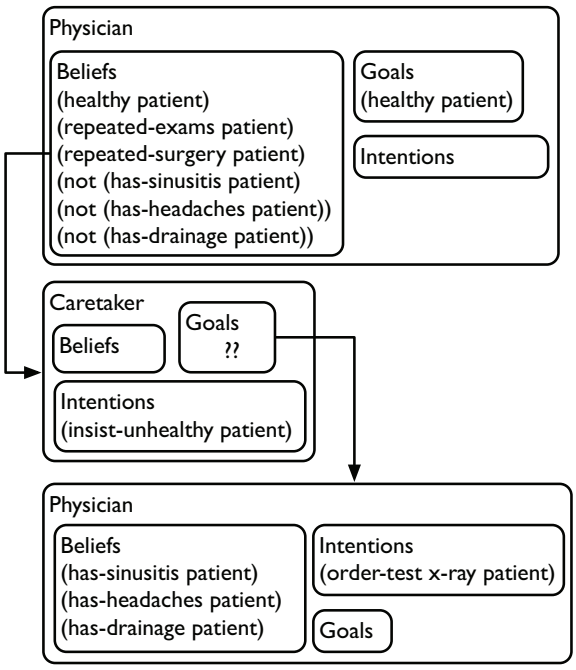


Figure 6: A snapshot of the physician's mental state while interacting with a caretaker who has Munchausen by proxy.

## 4 Deception, Ignorance, and False Beliefs

The scenarios in Section 3 dealt with deception, and one type of deception in particular. In each case, the patient was lying, which we earlier defined as involving the speaker's intent, false content, and the listener's response. However, other forms of deception, including bullshit and paltering also deserve attention from the cognitive systems community. According to Frankfurt (2005), bullshitters, unlike liars, are unconcerned with the veridicality of their statements and therefore have no intent to deceive. Instead, their intent is to hide their own disconcern for the truth. In contrast, paltering includes lies of omission, where the speaker leaves out the truth or asserts a partial truth (Schauer & Zeckhauser 2009). Additionally, details remain to be worked out in order to tease apart the interplay between the intent to lie (and deception, in general) and other mental states, such as false belief and ignorance.

To give an example of false belief, consider a modification to Scenario 1 where the drugs involved are the analgesics Advil and Motrin.

**Scenario 3**
*Physician*: "Tell me what's wrong."
*Patient*: "I have a terrible headache!"
*Physician*: "Have you tried pain relievers like Advil?"

*Patient*: "Oh no, doctor, I'm allergic to Advil."
*Patient*: "Last time, Motrin really helped."
*Physician*: ". . ."

As in the Tylenol–Vicodin case, the medications mentioned here share the same ingredient, this time ibuprofen, and the patient's claims are inconsistent. The fundamental difference is that, in this scenario, neither medication contains a narcotic or other recreational pharmaceutical. The lack of this additional factor prevents inference to an ulterior motive, and the physician is free to infer a false belief. For instance, the patient may have misremembered a prior allergic reaction or may have been confused about the drug names. Although the definitions of false belief and deception clearly differ, correctly distinguishing instances of each presents a difficult challenge for the research community.

Ignorance differs from false belief in that for some proposition $p$ that holds in the world, the agent with false belief will explicitly hold $\neg p$, whereas the ignorant agent will hold neither $p$ nor $\neg p$. In conversation, however, agents may make statements that they do not actually believe as way to cover their ignorance. At first, the response resembles bullshit, but unlike the bullshitter, the ignorant agent may value veridicality. Suppose a doctor asks, "Are you allergic to ibuprofen?" The patient may respond, "Yes," not because he knows what ibuprofen refers to and is aware of an allergy, but because he does not want to appear ignorant by asking for clarification. The content of the question suggests erring on the side of caution, so the patient provides a meek response.

Ignorance and false belief are superficially similar in that they may lead to the same inconsistent statements. Nevertheless, our framework can model them differently due to its ability to represent mental states. Still, dynamically differentiating between the two may require actions on the part of the agent, such as cross-checking the statement or asking more pointed questions. Handling this distinction as well as those amongst other forms of deception including bullshitting and paltering is a matter for future investigation.

## 5 Future Research and Conclusion

The computational framework that we introduced shows promise, but more research is required to meet our goal of a socially aware cognitive system. The obvious next steps are to implement the framework and to connect it to an inference system. Past experience in this direction has led us to conclude that the integration must be deep enough that the inference system can work directly with the memory partitions and navigate them as it reasons. The primary challenge here is the non-monotonic inheritance of mental content from one agent partition to another. A reasonable solution will involve finding the right balance between the validity of the ascriptions and computational complexity. We conjecture that an abductive reasoner, such as AbRA (Bridewell & Langley 2011), will cope better with this problem than deductive approaches which are known to be brittle when faced with inconsistency.

As we continue to develop and implement this framework, we also recognize the need to formalize its behavior. We would like to have a declarative theory of deception that supports detection and distinguishes it from false belief and ignorance. Sakama *et al.* (2010) have made some progress in this direction, but their approach is still limited to static definitions of deception. Additionally, Lee (1998) characterizes deception and false belief using an extended version of ViewGen, but he emphasizes ruling out these conditions in order to justify the inference of pragmatic implicatures in dialogs. Both models need substantial adaptations to treat the effects of a recognized (or suspected) lie on the mental state of an agent.

In this paper, we developed a framework for socially aware inference and introduced deception as a topic for cognitive systems research. Additionally, we identified naivete and skepticism as distinct agent attitudes toward deception plausibly encountered in realistic scenarios. Moreover, we demonstrated how models in our framework can represent these attitudes and dynamically adjust between them in response to deception. Finally, we briefly discussed the challenges involved in distinguishing deception from states producing similar behavior, such as false belief. The existence of examples such as Münchausen by proxy demonstrates that any effective, plausible framework for mind-reading cannot assume the veridicality of an agent's statements and, instead, must possess the ability to model and detect complex deceptive behavior.

## Acknowledgments

This research was funded by the Office of Naval Research under Contract No. ONR-N00014-09-1-1029 and a postdoctoral fellowship from the McDonnell Foundation Research Consortium on Causal Learning. We also thank Amar K. Das and Wei-Nchih Lee for helpful conversations.

## References

Alchourrón, C.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50, 510–530.

Awadallah, N.; Vaughan, A.; Franco, K.; Munir, F.; Sharaby, N.; and Goldfarb, J. 2005. Munchausen by proxy: a case, chart series, and literature review of older victims. *Child Abuse & Neglect*, 29:931–941.

Ballim, A., and Wilks, Y. 1991. *Artificial Believers*. Hillsdale, NJ: Lawerence Erlbaum Associates, Inc.

Baltag, A.; Moss, L.; and Solecki, S. 1998. The logic of common knowledge, public announcements, and private suspicions. In *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge*, 43–56. Evanston, IL

Bridewell, W., and Langley, P. 2011. A computational account of everyday abductive inference. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*. Forthcoming.

van Ditmarsch, H.; van der Hoek, W.; and Kooi, B. 2007 *Dynamic Epistemic Logic* Berlin, Germany: Springer.

Frankfurt, H. G. 2005. *On Bullshit*. Princeton, NJ: Princeton University Press.

Lee, M. 1998. Belief, Rationality, and Inference: a General Theory of Computational Pragmatics. Ph.D. diss., Dept. of Computer Science, The University of Sheffield, Sheffield, England.

Lenat, D. 1998. *The Dimensions of Context-Space*. `http://www.cyc.com/doc/context-space.pdf`.

Lewis, D. 1973. *Counterfactuals*. Malden, MA: Blackwell Publishers, Inc.

Rao, A. S., and Georgeff, M. P. 1995. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems*, 312–319. San Francisco, California.

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

Schauer, F., and Zeckhauser, R. 2009. Paltering. In *Deception: From Ancient Empires to Internet Dating*, 38–54. Stanford, CA: Stanford University Press.

Squires, J. E., and Squires, R. H., Jr. 2010. Munchausen syndrome by proxy: ongoing clinical challenges. *Journal of Pediatric Gastroenterology and Nutrition*, 51:248–253.

Wellman, H. M. 1990. *The Child's Theory of Mind*. Cambridge, MA: MIT Press.