Changing Minds by Reasoning About Belief Revision: A Challenge for Cognitive Systems

Will Bridewell WILL.BRIDEWELL@NRL.NAVY.MIL
Naval Research Laboratory, 4555 Overlook Ave. SW, Washington, DC 20375 USA

Paul Bello PAUL.BELLO@NAVY.MIL
Office of Naval Research, 875 N. Randolph St., Arlington, VA 22203 USA

Abstract

In this paper, we explore the representational and inferential requirements for supporting a rich notion of *belief revision*. Our analysis extends beyond the typical case of a single agent revising its beliefs in light of new information into the realm of social engagement. More to the point, we argue that although belief revision mechanisms surely operate at the level of single agents, we must also consider the need to *lift* an agent's understanding of the belief revision process to the knowledge level in order to intentionally guide other agents' revision processes with whom it socially interacts. In exploring belief revision at the knowledge level, we identify reasons for rejecting classical formulations of the problem and identify constraints by which alternative accounts must abide.

1. Introduction

Belief revision is a common result of human dialogue and the reason for many conversations. People chat about the world, learn new facts, and give up outdated or incorrect beliefs. Sometimes this process happens without obvious results, but other times we see it played out in arguments and discussions. Clearly, intelligent agents must be able to update their own beliefs and knowledge. Traditionally, artificial intelligence (AI) has taken a limited view of belief revision, seeing it as an automatic, formalized means for truth maintenance. Here, we claim that as researchers move toward modeling socially aware cognitive systems, they must also change their view of belief revision—its purpose, its operation, and its flexibility.

Historically, AI inherited its view on belief revision from logicians, who strongly emphasize the importance of consistency (Alchourrón, Gärdenfors, & Makinson, 1985). In this work, beliefs are typically seen as elements in a theory about the world that may be expanded, contracted, and when consistency is threatened by a new belief, revised. Carrying out revision involves removing the set of elements from the theory that imply a conflict with a new belief. Since there may be several such sets, the standard approach involves an appeal to epistemic entrenchment (Gärdenfors & Makinson, 1988), which orders axioms based on logical entailment. Loosely stated, the general principle is to be conservative, removing as few beliefs as possible to maintain consistency. The prominence given to not only consistency but also automaticity is shared by many systems in AI, including varieties of truth maintenance systems (McAllester, 1990; de Kleer, 1986).

Our operative assumption for this paper is that agents do not come equipped with belief revision mechanisms that operate automatically under the hard constraint of preserving global consistency. To illustrate our position, we discuss two scenarios. The first draws attention to a person who simultaneously believes a statement and its direct negation. The second uses a social interaction called *impression management* to illustrate the need for a richer approach not only to beliefs but also to belief revision. Using these examples, we will argue that (1) beliefs per se are more complex than their treatment in the belief revision and AI literatures suggests, (2) global and even contextualized consistency constraints on beliefs may be overly optimistic representations of real believers, and (3) belief revision is neither atomic nor automatic but an interruptible and manipulable process. The latter will be shown by constructing and exploring an example of impression management cast as information-processing that involves mental state ascription between agents—often called mindreading (Goldman, 2006).

After stating our case, we discuss belief revision strategies that an agent can use to change another's mind. As complex as these strategies might seem, in isolation they do not adequately capture the kind of reasoning required to effectively engage in argument, deceit, or other activities meant to manipulate the beliefs of others. To this end, we argue by example that reasoning about strategies and combinations of strategies via mindreading is necessary to achieve one's desired social goals.

To be clear at the outset, we are not suggesting or developing a formal or computational model of the phenomena discussed in this paper. Our goal is to highlight the richness of beliefs and belief revision as viewed through the lens of everyday social activity. We offer this analysis to encourage formalization and eventual implementation of rich, social models of belief revision. The complexity revealed by the examples that we discuss suggests that more than one computationally oriented paper is necessary to do full justice to the topic. We end by commenting on the constraints that socially motivated belief revision places on computational systems, noting that these will be suggestive and not prescriptive.

2. Detecting and Responding to Inconsistency

In this section, we examine the nature of belief and the consistency of beliefs within a single agent. For several decades, consistency and closure under deduction have been hallmarks of the belief revision and epistemic reasoning frameworks developed by AI researchers and philosophers (Hintikka, 1962). The AGM postulates mentioned in the introduction are typically expressed as conditions under which a system preserves or fails to preserve consistency after adding a new candidate belief to an agent's storehouse of existing beliefs (Koons, 2013). Rather than being descriptive, the AGM postulates and axioms for formal logics of belief are prescriptive, defining normative conditions for rationally revising one's beliefs in light of new information (Stalnaker, 1993). Standing in sharp contrast, cognitive systems must negotiate the complexity of real-world social situations and thus represent what appears to be the decidedly non-normative epistemic practices of real human beings.

Consider the following paraphrased example adapted from (Schwitzgebel, 2010):

> *Juliet the implicit racist.* "Many Caucasians in academia profess that all races are of equal intelligence. Juliet, let's suppose, is one such person, a Caucasian-American

> philosophy professor. She has, perhaps, studied the matter more than most: She has critically examined the literature on racial differences in intelligence, and she finds the case for racial equality compelling. She is prepared to argue coherently, sincerely, and vehemently for equality of intelligence and has argued the point repeatedly in the past. And yet Juliet is systematically racist in most of her spontaneous reactions, her unguarded behavior, and her judgments about particular cases. When she gazes out on class the first day of each term, she can't help but think that some students look brighter than others—and to her, the black students never look bright. Juliet could even be perfectly aware of these facts about herself. She could aspire to reform. Self-deception could be largely absent. We can imagine that sometimes Juliet deliberately strives to overcome her bias in particular cases. She sometimes tries to interpret black students' comments especially generously, but it is impossible to constantly maintain such self-conscious vigilance."

From the perspective of an outside observer, Juliet holds inconsistent beliefs. Her all-things-considered convictions come apart from whatever kind of belief-like mental content guides her unguarded actions. In a sense, she believes both that different races possess equal intelligence and that they do not. This is a notably different story from the one told about rational agency in the AI and philosophy literature.

When considering suitably realistic scenarios such as the one above, a tidy picture of belief as a simple predicate or logical operator is ineffective. Furthermore, representing beliefs as probability distributions also fails to solve the problem.[1] Juliet is in no way uncertain about her egalitarian beliefs. Instead, we can separate her beliefs into implicit action-guiding mental content and explicit representations that she can reflect reflect upon. Even when Juliet reflects on her behavior and presumably self-ascribes the belief "I sometimes act like a bigot," she neither gives up her prior egalitarian beliefs, nor takes extra care to align her actions with her professed belief in racial equality. This is a poignant example of how beliefs are contextualized rather than stored in a deductively closed, global belief-box that enforces internal consistency. Egan (2008) provides an extended argument for the benefits of fragmented belief-bases over the globally consistent alternative.

We take our argument one step further and claim that even the notion of an automatically maintained local consistency is implausible. In the context of this example, Juliet can think of herself separately as "Juliet the racially biased" and "Juliet the racially blind" by establishing two separate, locally consistent versions of herself and thinking of each version objectively. However, unless we adopt an ad hoc notion of a fragmented self, Juliet is and acts as a single person who holds both beliefs about racial equality simultaneously. The literature on cognitive dissonance (Harmon-Jones & Mills, 1999) rests itself on this capacity for self contradiction however uncomfortable it makes us

[1] The relationship between full and partial beliefs occupies the time of many formal epistemologists. Paradoxes of rationality such as the Lottery Paradox (Kyburg, 1961) and the Paradox of the Preface (Makinson, 1965) illuminate the difficulties faced by giving a formal account of full belief in terms of partial belief (although see Pollock (2008) for a notable, computational effort). Much current thinking on these matters relies on suspending judgment in certain circumstances until an adequate evidential threshold is crossed. It seems clear that simply treating beliefs as subjective probabilities will not bridge the divide between full and partial belief; however, it may be possible to do so by exposing the choice and strength of thresholding to the knowledge level, consistent with the rest of our story about knowledge-driven revision.

feel. Moreover, the dissonant experience persists in time and guides actions intended to resolve the contradiction. It is this idea of belief revision as a process, one that is manipulable both by an agent herself and by external agents, to which we now turn.

3. Impression Management: an Example

As a prime example of a common and particularly intricate social process, consider *impression management* (Malle, 2004). Roughly speaking, impression management takes the form of an agent talking and acting in ways that signal another agent to ascribe specific traits, beliefs, or inclinations to him. Importantly, the agent's behavior is intentional and may not reflect his actual mental state. To illustrate, we use a fictional scenario from an office environment.

> Bob works for Jane at a Silicon Valley technology company. Recently Jane stated that user studies have proven that the new logo for the company's software should be a mantis shrimp. In addition, she would like her entire team on board with this decision. Due to an overheard conversation where Bob said, "I couldn't care less about the logo," she believes him to be disinterested. Through the corporate grapevine, word about her impression got back to Bob who was relieved to find out that Jane is unaware of his intense distaste for her decision. Personally, he finds the proposal ridiculous, preferring the honey badger. Nevertheless, Bob enjoys his job and would rather not risk it by attempting to talk Jane out of the new logo. Instead, he would rather give Jane the impression that he is at least open to the mantis shrimp as a possibility.

To meet his goal, Bob needs to manipulate Jane's belief revision process. In a conversation with Jane about the relative merits of her design, Bob might assert his original disinterest followed by a series of statements indicating a growing fondness for the logo. Crucially, Bob would still consider the proposal ludicrous, but would have (potentially) managed to change Jane's impression of his mental content by making laudatory comments about the logo and acting in corresponding ways.

4. A Mindreading Framework

To understand this example, we need a way to characterize Bob's mental state, including his beliefs for and goals about Jane's mental state. To this end, we adopt the Framework for Identifying Deceptive Entities (FIDE) (Isaac & Bridewell, in press). This framework depicts mental content about an agent's self and others, known ignorance, and ulterior motives. Notably, FIDE models are synchronic and make no commitment to dynamic transitions between mental states. Currently, the formalism supports default ascription and the overriding of standing norms, which we describe below, but lacks the axioms that doxastic logics (Hintikka, 1962) or the BEL operator (Allen, 1995) provide. FIDE also lacks computational mechanisms for simulating social behavior (Stuhlmüller & Goodman, in press; Pynadath & Marsella, 2005), but we discuss system requirements Section 7.

A FIDE agent consists of a set of agent models with one designated as the self. In discussing the example from Section 3, we take Bob's perspective, creating a self-designated model for him and a separate one for Jane. Both models store propositional attitudes that correspond to an agent's mental

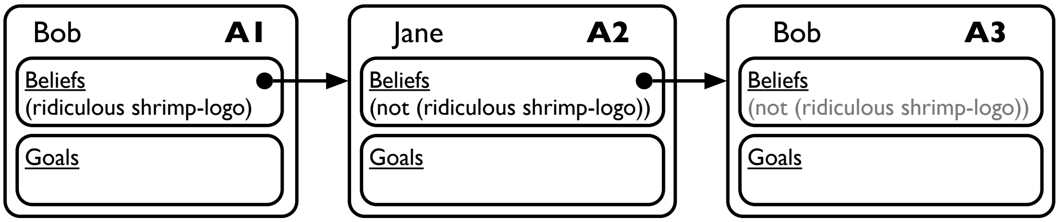

Figure 1. Default ascription of a belief in FIDE. A1, A2, and A3 are agent models. Here, Bob believes (a) that the shrimp logo is ridiculous and (b) that Jane believes that it is not ridiculous. Also, through default ascription, Jane's belief is available to the nested model of Bob (A3). The lighter shade indicates that the belief is not explicitly represented.

content. For instance, Bob believes that the proposed logo is ridiculous, represented syntactically as `(B bob (ridiculous shrimp-logo))`, and Bob has a situationally relevant maintenance goal to stay employed, written as `(G bob (employed bob))`. Additionally, FIDE supports *standing norms*,[2] which include general behavioral guidance like the Gricean maxims "be truthful" or "be relevant." Even though we refer to these as norms, we assume that agents maintain standing goals to abide by them, subject to the demands imposed by resources and the agent's current situation.[3] We assume that all agents share these goals and, importantly, that they can defeat them with a higher priority, ordinary goal. Returning to Bob, we find that he values honesty in social interactions, but his goal to stay employed defeats the standing norm T (be truthful) in his current situation.

Within an agent, models connect to each other through modalities, which enables mindreading. In this sense, FIDE supports syntactic statements like `(B bob (B jane (B bob (not (interesting shrimp-logo)))))`, Bob's belief that Jane thinks he finds the logo uninteresting, and `(B bob (G jane (B bob (interesting shrimp-logo))))`, his belief that Jane would like him to be interested in the logo. To avoid an excessive representational burden, we adopt ideas from ViewGen (Ballim & Wilks, 1991) and Polyscheme (Bello, 2012) that support *default ascription* from a parent agent to a child agent (e.g., from Bob to Jane). Figure 1 uses the example from Section 3 to illustrate this idea. According to default ascription, in the absence of contradictory evidence, the model asserts that Bob believes that Jane believes that the mantis-shrimp logo is ridiculous. To block this ascription, the model needs the proposition `(B bob (B jane (not (ridiculous shrimp-logo))))`. Again, by default ascription, indicated in the figures by a light-gray font, this belief is accessible in Bob's model of Jane's model of Bob.

In addition to the default ascription of beliefs, we also take the stance that agents generally ascribe their own rules and mechanisms of inference to others. Similarly, we hold that these elements can be overridden, as occurs with an interaction between an expert and a novice, although this is not

[2] Standing norms are related to the *ethical goals* described in (Isaac & Bridewell, in press). We have widened the scope of these general constraints on behavior to include epistemic goals such as "be conservative in revising existing beliefs," or "be rational." These do not intrinsically have any ethical character in the way that the Gricean maxim of quality (i.e., "be truthful") might.

[3] The explication of standing norms as a type of goal suggests that we are conflating obligations and goals, which clearly are separate propositional attitudes. Technically speaking, what we call *standing norms* are goals whose content is a complex conditional of the form "If possible and if there exists no defeating reason, then act in accordance with *N*," where *N* is a definite description of the norm to be abided by.

yet supported in FIDE. Ascribing the same mechanisms across agents lets us treat social reasoning as mental simulation. That is, once we connect a model of Jane to the self-designated agent, Bob, he can use his own inference system to power reasoning from Jane's perspective. Furthermore, inasmuch as the self-designated agent can consider its own reasoning processes, it can also apply that knowledge to understanding—and manipulating—the reasoning of another agent.

As a final feature, FIDE supports the representation of *ulterior motives*. Recall that when introducing standing norms, we mentioned that Bob's situational goal to stay employed defeats his ethical goal T. In this case, we can state Bob's ulterior motive as `(G bob (employed bob))` and include a meta-level goal `(G bob (defeat '(employed bob) T))`. The quote operator identifies an argument as a *proposition identifier* as opposed to propositional content. The proposition `(employed bob)` is named by the quoted, string representation of its content. We say that the "defeat" predicate indicates a meta-level goal because it describes the relationship between two normal goals. Typically, we use the phrase "ulterior motive" to describe a hidden agenda, but within FIDE the term applies specifically to those goals that defeat standing norms. As the next section will show, ulterior motives play a central role when manipulating another agent's belief revision process.

5. Impression Management in FIDE

Although FIDE is principally a representational framework, it provides a foothold for discussing mental states in complex, social situations. Specifically, when illustrating the impression management example, we can view the mental content of the agents both before and after a successful (from Bob's perspective) conversation. This representation provides a perspective on social reasoning that offers two benefits. First, the model makes explicit the agent-level distinctions necessary for routine interaction. These elements include the beliefs, goals, and motives of oneself and others. Second, the mental states reveal the requirements of a cognitive system capable of routine, but complex, social engagement. In particular, such a system would need the ability to reason deliberately about the consistency of its own mental states, the consistency of other agents' mental states, and the mental processes that influence belief revision.

The scenario from Section 3 exhibits several characteristics that point toward a richer theory of belief revision than traditionally expressed in the AI literature. Specifically, Bob must

1. maintain a model of Jane's mental state to recognize his desire for revision;
2. override a standing norm to be truthful in pursuit of a highly prioritized goal; and
3. reason about Jane's belief revision process to affect the outcome.

All these capabilities must respond to the changing conditions brought about during Bob's conversation with Jane. After applying FIDE to this example, we discuss our claim that belief revision is temporally extended and directly manipulable.

5.1 Representing Bob's Mental State

Figure 2 illustrates key elements of Bob's mental state before he speaks with Jane as described in the example from Section 3. As shown in agent model A1, Bob personally finds the mantis-shrimp logo to be ridiculous and has a goal to stay employed that defeats T. His model of Jane, A2, contains

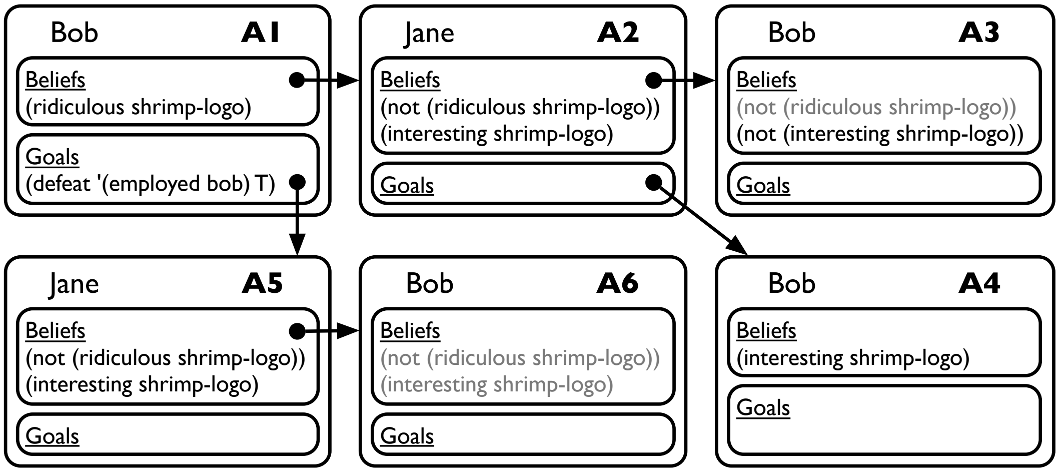
Bob A1
Beliefs
(ridiculous shrimp-logo)
Goals
(defeat '(employed bob) T)
Jane A2
Beliefs
(not (ridiculous shrimp-logo))
(interesting shrimp-logo)
Goals
Bob A3
Beliefs
(not (ridiculous shrimp-logo))
(not (interesting shrimp-logo))
Goals
Jane A5
Beliefs
(not (ridiculous shrimp-logo))
(interesting shrimp-logo)
Goals
Bob A6
Beliefs
(not (ridiculous shrimp-logo))
(interesting shrimp-logo)
Goals
Bob A4
Beliefs
(interesting shrimp-logo)
Goals

Figure 2. Before impression management. A1–A6 are agent models. Before talking to Jane, Bob believes (a) that the shrimp logo is ridiculous, (b) that Jane believes that it is not ridiculous, and (c) that Jane thinks that Bob finds it uninteresting. Jane has a goal for Bob to be interested in the logo, and Bob has a goal for Jane to believe that he is. T represents the standing norm to be truthful.

both a positive belief that the logo is interesting and a negative one that blocks the default ascription of Bob's assessment.

From Bob's viewpoint, Jane holds two models of him, a model of his beliefs, A3, and a model of her goals for his beliefs, A4. The content of these models are contradictory, but the models themselves are internally consistent. If A3 and A4 are accurate, Jane will attempt to alter Bob's beliefs, possibly by saying that the logo is interesting and arguing for him to adopt the belief `(interesting shrimp-logo)` and to refute the contradicting one. Bob, of course, has other plans. His goal is to have Jane, A5, believe `(interesting shrimp-logo)` while covertly maintaining `(ridiculous shrimp-logo)`. That is, instead of changing his mind, he can act as if he has changed his mind.

Figure 3 shows the aftereffects of a successful conversation between Bob and Jane. Here, default ascription states that A3 is consistent with A2 (i.e., A3 can access all beliefs in A2). Notably both Bob and Jane achieved their goals (assuming that Bob's mental models are accurate) even though their individual beliefs conflict.

5.2 A Cognitive Systems View of Belief Revision

We began this section with a list of requirements for belief revision, starting with the ability to model another agent's mental state and recognize one's own desire to revise it. As Figure 2 shows, a framework like FIDE can represent other agents, but a cognitive system needs more than this to recognize a need to revise another agent's beliefs. The minimal level of support involves the identification of a goal state that diverges from the perceived state of the world. However, the emphasis on mental attitudes distinguishes this case from normal goal-directed planning scenarios. For instance, Bob could align his belief model of Jane with his goal model without taking any

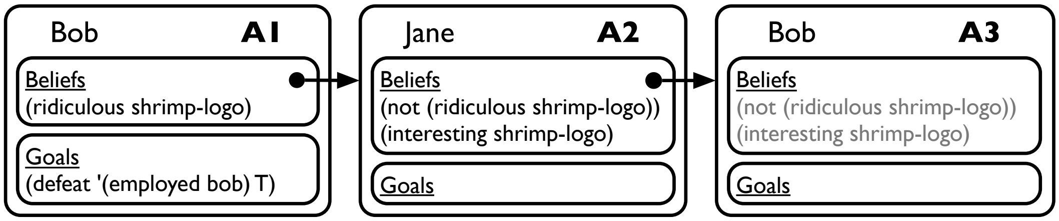
Bob
A1
Beliefs
(ridiculous shrimp-logo)
Goals
(defeat '(employed bob) T)
Jane
A2
Beliefs
(not (ridiculous shrimp-logo))
(interesting shrimp-logo)
Goals
Bob
A3
Beliefs
(not (ridiculous shrimp-logo))
(interesting shrimp-logo)
Goals

Figure 3. After impression management. A1–A3 are agent models. After successfully manipulating Jane's beliefs, Bob believes (a) that the shrimp logo is ridiculous, (b) that Jane believes that it is not ridiculous and is interesting, and (c) that Jane thinks that Bob agrees. T represents the truthfulness standing norm.

external action (e.g., through rationalization, "Jane doesn't really care what I think."). Determining the need to revise another's beliefs involves a value judgment that scopes over outcomes. In any scenario like this one, an agent may (a) alter its own beliefs about the other agent, (b) attempt to alter the other agent's beliefs, or (c) prefer the unsatisfied goal to explicit action.

Assuming an agent decides to revise another's beliefs, it may need to violate a standing norm. This capability is reflected in the current example where Bob deems truthfulness less important than continued employment: `(defeat '(employed bob) T)`. By deciding to violate the norm, Bob increases the number of strategies available for altering Jane's beliefs. More generally, an agent's normative commitments block some avenues of action.

One might argue that a system that cannot lie has no need for standing norms or the ability to defeat them. For example, with respect to norm T, we as responsible researchers should build veridicality into any cognitive system. Certainly this is the typical approach in AI. However, there are times when lying is the right thing to do. For instance, protecting organizational secrets from hostile parties may require more than truthful misdirection, and social interactions frequently require some form of deception to maintain congeniality.[4] Moreover, agents that interact with humans will need to reason about their trustworthiness and the effects of lying. However, once an ethical boundary can be crossed, provisions must be made to ensure that it is not always crossed. Therefore a cognitive system needs the capacity to both represent and override standing norms of the kind we have discussed above.

Even in the absence of deception, revising one's beliefs about another agent by altering their beliefs requires more than extended modeling capabilities. Suppose Bob and Jane have started their dialog, and Bob says, "You know, Jane, the mantis-shrimp logo is really interesting." Whether Jane immediately accepts Bob's statement or cordons it off for potential later acceptance, eventually she will need to resolve a contradiction. Standard approaches to belief revision would isolate the process, having Jane update her beliefs automatically and immediately. This characterization leaves Bob and agents like him unable to predictably manipulate others' beliefs and with no reason to note the contradiction at all. Instead, we claim that such cognitive systems must lift consistency constraints from the architectural level to the knowledge level, using this knowledge to recognize contradictions as opportunities to influence belief revision.

[4] On this latter point, Hollywood has explored the pathological implications of unmitigated truthfulness in films such as *Liar Liar* (1997) and *The Invention of Lying* (2009) to comic effect.

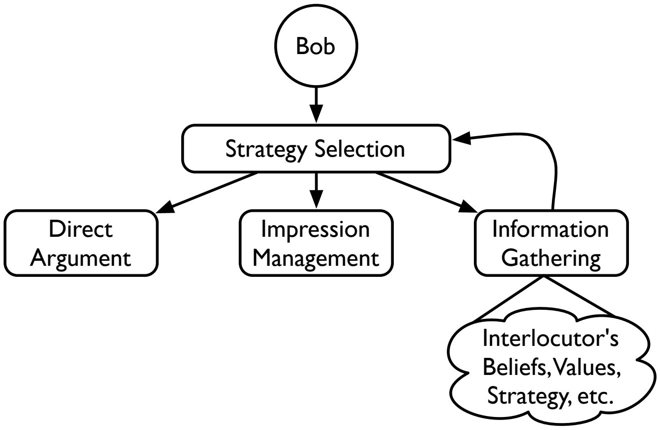

Figure 4. Strategy selection. When seeking to revise another agent's beliefs, Bob can engage in any one of several strategies, including information gathering whose results feed back into the selection mechanism.

This claim implies the ability to entertain inconsistent beliefs without automatically importing all deductively valid inferences. Instead, a cognitive system must approach an inconsistency directly, identifying a strategy either to reconcile beliefs or to preserve the ambiguity. Importantly, the strategy is not likely one of immediate and automatic belief revision as in the AGM tradition, but one of gathering evidence either through engaging with the world or through deliberative reasoning about the contradictory thoughts. That the inconsistency remains over time and that the mental mechanisms for handling the inconsistency take place in the context of a perceiving and acting system gives other agents the opportunity to interject their own preferences into the process. Bob can guide Jane's thoughts. With this in mind, we next delve deeper into the social interactions associated with belief revision.

6. Alternative Belief-Revision Strategies

Impression management is just one strategy among several for manipulating belief revision through dialog. To better understand richness of potential interactions, we look at a strategy that maintains truthfulness but violates another standing norm. We also examine ways in which an agent might combine multiple strategies. Then we look at how an agent might gather information to help select among the many possible strategies.

6.1 Direct Argument

Figure 4 introduces *direct argument* as one alternative. If Bob were to follow this strategy, he might straightforwardly attempt to convince Jane that the mantis-shrimp logo is ridiculous. In this case, his dialog moves could include highlighting unintended relationships suggested by the logo (e.g., that a competitor's logo is an octopus, a natural predator of the mantis shrimp). He could also emphasize salient features of the honey badger, his preferred alternative.

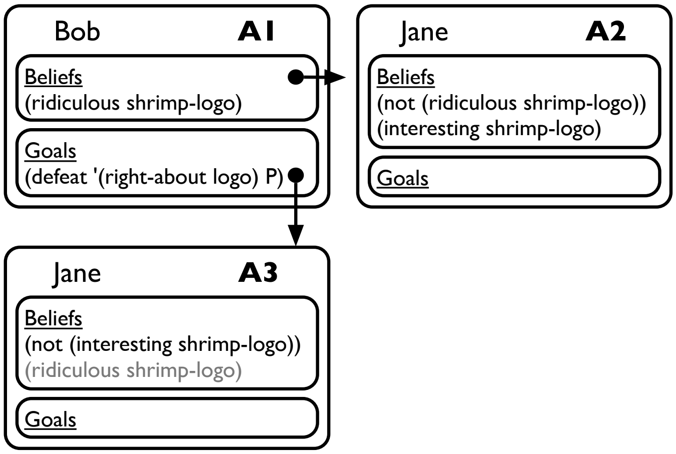
Bob
A1
Beliefs
(ridiculous shrimp-logo)
Goals
(defeat '(right-about logo) P)
Jane
A2
Beliefs
(not (ridiculous shrimp-logo))
(interesting shrimp-logo)
Goals
Jane
A3
Beliefs
(not (interesting shrimp-logo))
(ridiculous shrimp-logo)
Goals

Figure 5. Direct argument. A1–A3 are agent models. Bob's beliefs shown in A1 and A2 are identical to those in Figure 2, but his goal now here is to convince Jane (A3) that the mantis shrimp logo is not interesting. P represents a standing norm for politeness.

More abstractly, direct argument presumes that `(B agent1 Q)` and `(B agent1 (B agent2 (not Q))` for any two agents and some proposition Q. Figure 5 reflects this pattern in the context of Bob and Jane. Bob believes that the shrimp logo is ridiculous and that Jane believes the shrimp logo is not ridiculous. However, this condition alone is insufficient; Bob must also want Jane to believe differently. Referring back to Figure 5, we see `(G bob (B jane (not (interesting shrimp-logo))))` and `(G bob (B jane (ridiculous shrimp-logo)))`, by default ascription, in A3. The important point here being the ability to distinguish between what Bob *thinks* Jane believes (A2) and what he *wants* her to believe (A3).

Even so, a direct contrast between beliefs and goals is not the only precondition for a direct argument. For instance, Bob could have a goal for Jane's beliefs but never act on it. To get a sense for what is missing, recall from Section 5.2 that impression management involves an ulterior motive that defeats the Gricean maxim of quality. For example, the ulterior motive of staying employed led to Bob's deception through impression management in Section 3. In comparison, direct argument *simpliciter* involves violating a standing norm prescribing, "One ought to be polite," denoted as P in Figure 5. All that we mean by politeness here is that agents ought to be non-confrontational by default, and perhaps especially so in situations where social roles determine the degree to which a speaker can be confrontational with a given hearer. Overriding this norm relaxes constraints on the kinds of utterances an agent is apt to produce, opening the door for direct argument and other strategies. For the purpose of brevity we refrain from exploring possible drivers for being impolite, but note that existing work by Briggs and Scheutz (2013) on modeling socially driven politeness-constraints for utterance generation seems highly relevant.

6.2 Combining Approaches to Belief Revision

So far, we have considered strategies for belief revision in isolation, but real interactions often involve nesting strategies in arbitrary ways. To illustrate more complex instances of social belief-revision, consider the following variation on the Bob and Jane vignette (with the original portion shown in italics):

> *Bob works for Jane at a Silicon Valley technology company. Recently Jane stated that user studies have proven that the new logo for the company's software should be a mantis shrimp. In addition, she would like her entire team on board with this decision. Due to an overheard conversation where Bob said, "I couldn't care less about the logo," she believes him to be disinterested. Through the corporate grapevine, word about her impression got back to Bob who was relieved to find out that Jane is unaware of his intense distaste for her decision. Personally, he finds the proposal ridiculous, preferring the honey badger.*
>
> *Nevertheless, Bob enjoys his job and would rather not risk it by attempting to talk Jane out of the new logo.* However, he has the impression that Jane finds him to be rather milquetoast and considers him neither a confidant nor a "company man" in the usual sense. Bob also knows that Jane rose to her position of relative authority in the company by being a keen judge of character and decidedly manipulative in her own right. He gets the sense that the vanilla-flavored pandering that often accompanies impression management is too much to risk given their lukewarm relationship and her uncanny knack for sniffing out sycophants. Instead, he engages in a bit of reverse psychology, stating up front to Jane that he deeply dislikes the mantis-shrimp logo. He avoids arguing the point and explicitly states that he wants only to be honest with her.
>
> For her part, Jane is taken aback by Bob's willingness to be both opinionated and forthcoming about it. Such behavior violates her expectations about his past lack of initiative. She welcomes his newly demonstrated straightforwardness, tells him that he is entitled to his opinion, and asks him if he would be willing to nominally advocate for the mantis-shrimp logo when the matter was put to the team for committee vote. He happily agrees to do so.

The interaction described above is what you might expect out of typical office or faculty meetings. Here, Bob employs a strategy undiscussed so far in this paper: merely being honest. In this case, Bob does have an ulterior motive. He wants to curry favor with the boss and knows that the easiest way to do so is by portraying himself as a company man who also has strength of conviction. In this sense, being honest can be a form of impression management. Bob is not trying to manage Jane's impression of what he thinks about a particular issue—rather, he is interested in making Jane think that he cares more than he actually does. He wants to manage her impression of his traits. Their agreed-upon arrangement comes along with an obligation for Bob to facilitate the direct argument that Jane intends to make to the team by managing their impression of his attitude toward the mantis-shrimp logo.

Similar examples can also involve the use of impression management inside of a direct argument. Suppose Bob decides to directly argue for the honey-badger logo and against the mantis shrimp. Jane's tone of voice, facial expressions and body language may seem to indicate that he has taken the argument too far. One way Bob might rescue himself is by voicing uncertainty about the strength of his reasons for the part of the argument that riled Jane. In fact, he can "walk back" any number of statements if he gets a clear sense that he is pushing too hard. In any event, the space of basic belief-revision strategies is likely not exhausted by direct argument and impression-management. We leave the discussion of other strategies to the future along with whether they are arbitrarily embeddable within one another.

6.3 Choosing Among Strategies by Gathering Information

In addition to implementing a strategy, agents may interleave *information gathering* activities to refine their strategies or to determine which particular strategy to apply. Deciding among direct argument or impression management will largely be a function of gauging how strongly agents hold certain target beliefs and goals. To this end, an agent will engage in information gathering, which in and of itself does not result in manipulating the revision process of another agent. Instead, information-gathering attempts by agent A directed at agent B may in some cases be aimed at eliciting knowledge about the level to which B's beliefs about the content under discussion are epistemically entrenched. Analogously, information gathering can operate separately or in parallel to estimate the strength of B's resolve with respect to its goals and intentions for the interaction.

For agent A to lead B toward a desired conclusion, A must be able to reason about the belief revision process. So, as with consistency constraints, we must lift the rules governing belief revision from the architectural level to the knowledge level. Returning to Bob and Jane, through argument or cajolery, Bob can sway Jane toward the beliefs that would benefit him. For instance, he may say, "The logo doesn't matter much to me," bolstering Jane's impression of his indifference. He could then follow this with, "But tell me about the benefits," eliciting Jane's argument. As she engages in what she might interpret as an attempt to revise Bob's beliefs, he collects information that allows him to hone his estimates of Jane's beliefs and goals—both in terms of their content and the relative strength with which she holds them. Armed with this information, Bob is in better position to select among different strategies to affect Jane's belief revision process. If he decides that he can safely get away with managing her impression of him, he could express growing interest, until finally saying, "I see what you mean, the mantis shrimp truly does offer the best representation of our product." However, if he gets the sense that she is still somewhat conflicted about the logo, he could engage in direct argument to lead her to see the benefits of the honey-badger logo that he prefers.

This particularly intricate process contains the following activity. Bob recognizes Jane's belief about and goal for his mental state. He knows that to meet her goal, she will expect to engage Bob's belief revision process. In response, Bob establishes an opportunity for Jane to make an argument by first creating common ground on his public opinion. Finally, Bob indicates an inauthentic change in belief, producing a mental state like the one shown in Figure 3. Reflecting on this example, we find it difficult to determine how any cognitive system could engage in impression management without the ability to reason about not only the content but also the processes in another agent's mind. This is illustrated clearly in the elaborated Bob–Jane vignette presented in the previous section. Bob knows that Jane has a talent for uncovering the hidden agendas of panderers and uses this information to help him choose a strategy. For this kind of complex reasoning to take place, Bob needs to simultaneously reason in a strategy space alongside the garden-variety, mental-state reasoning that we have discussed throughout the paper.

7. Final Remarks

In summary, we have argued for a more nuanced account of belief and belief revision motivated in part by the demands of our social environment. We have shown that impression management is an indispensable tool for navigating the social world, but one that comes with the burden of having to reason explicitly about inconsistencies in both our own set of beliefs, and those in the set of beliefs we deem our interaction partners to have. Along the way, we have encountered examples that suggest a list of desiderata that a cognitive system must satisfy in order to behave reasonably in complex social situations. While likely not exhaustive in scope, the following list can be generated by way of our analysis of Juliet's conundrum and the Bob–Jane interactions presented throughout the paper:

1. a framework for mindreading that supports the representation of nested mental structures;
2. methods of reasoning directly about inconsistency without terminating processing;
3. a sophisticated representation of belief as a mental state;
4. an abductive reasoning capability;
5. a representation of standing norms as constraints on action;
6. an associated defeasible reasoning capability that enables the re-ordering and violation of standing norms;
7. a representation of belief-revision strategies as elements in both an object language and a meta-language; and
8. a practical reasoning capability that operates over both object- and meta-level representations of strategies.

Requirement 1 follows directly from the nature of socially aware inference. When reasoning about the belief-revision processes of others, a cognitive system must, at a minimum, be able to entertain beliefs about their beliefs. Any movement at all toward more sophisticated scenarios involving interlocutors that have beliefs and goals for the system's beliefs and goals clearly requires a non-trivial degree of nesting. Committing to nested content makes it natural to separate beliefs into different, partially quarantined mental spaces. In fact, most of the figures scattered throughout the paper have adopted this convention.

Bundling mental states in this way lets an agent reason about either itself or its interlocutor as having mental states that they do not yet have. Having goals for mental states in this way facilitates selecting among and executing strategies to move an agent's current mental state toward a particular goal configuration, as suggested by Requirement 8. Bundling also provides a natural device for representing inconsistency. So long as we can reference the names of the bundles within our object-language, we can contextualize beliefs and define inconsistency at the level of bundles rather than within them, generating an expression of the form:

$$\text{In}(\phi, \text{A1}) \wedge \text{In}(\neg\phi, \text{A2}) \wedge \text{Agent}(\text{A1}, \textit{agent}) \wedge \text{Agent}(\text{A2}, \textit{agent}) \rightarrow \text{Inconsistent}(\textit{agent}, \phi, \neg\phi).$$

Where $agent$ is modeled as A1 and A2, and the relation "In" indicates that a proposition exists within a particular model. The consequent asserts that $agent$ holds inconsistent beliefs about ϕ.

While beyond the scope of FIDE's current representational commitments, one can also use contextualization to capture the differences between implicit and explicit beliefs. The crucial fragment of reasoning in the Juliet vignette is computationally modeled by Bello (2013) as a preliminary attempt to develop a more sophisticated model of belief as an attitude, in line with Requirement 3. The approach that Bello advocates partially addresses some of the counterintuitive consequences of adopting standard modal frameworks for representing and reasoning about beliefs as given by Hintikka (1962) and by Allen (1995); most especially those related to *logical omniscience*, which demands that rational agents believe each member of the set generated by closing belief under implication. Other approaches for avoiding omniscience can be found in the literature, but each fails to reasonably capture human-like believing for reasons beyond the scope of our discussion (Levesque, 1984; Lakemeyer, 1986; Konolige, 1986).

Requirement 4 concerns abduction, which is necessary for any model of mindreading that can take observations of overt behavior and infer the most likely set of mental states involved with their generation (Bello, 2013; Bridewell & Langley, 2011). Abductive reasoning is central to properly capturing the dynamics of the Juliet example, but also widely useful for populating bundles with mental-state information. Impression management, direct argument and other strategies for manipulating the belief revision process of other agents seem to universally involve suspending or violating a standing conversational norm such as the injunctions to be truthful and polite.

To this effect, we require a representation of conversational norms and a defeasible inference mechanism to modify the set of constraints on action that each norm imposes. Interestingly, we also need to be able to reason directly about the norms themselves. We saw this on display in the elaborated case of Bob and Jane, where Bob decided it was better to be truthful than to engage in the risky business of trying to manage Jane's perception of his beliefs.

Finally, our discussion of information gathering and extended interaction suggests that we need a higher-order language to represent strategies as first-class objects, so that they can be used in practical reasoning. Examples of formalisms where this could be captured include FOL (Weyhrauch, 1980) and CycL (Matuszek et al., 2006).

Admittedly, this is a very premature exploration of a rich and complicated matter. We have not provided a working computational example by way of simulation, nor have we attempted to write down a knowledge-level characterization of the belief-revision process. In essence, we have not axiomatized the human folk-concept of belief revision that gets used whenever we engage in impression management. Such an undertaking would be massive, requiring an unachieved characterization of the folk concept of belief itself, along with ancillary characterizations of choice, evidence, and other complex concepts. Nevertheless, even in our state of relative ignorance, it is important to continue to identify subtleties in conceptual definitions that challenge conventional wisdom whenever we find them—belief revision being no exception to this rule. And although further subtleties await discovery and explication, we hope that our contribution in this paper invites deeper consideration of the cognitive architecture underwriting such an important part of our lives as socially aware agents.

Acknowledgments

We give special thanks to Alistair Isaac for his role in developing the FIDE framework and for comments that improved the organization and content of this paper.

References

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

Allen, J. (1995). *Natural language understanding*. Redwood City, CA: Benjamin Cummins. 2nd edition.

Ballim, A., & Wilks, Y. (1991). *Artificial believers*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bello, P. (2012). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems*, *1*, 59–72.

Bello, P. (2013). Folk concepts and cognitive architecture: Mental simulation and dispositionalism about belief. *Proceedings of the 2013 Meeting of the International Association for Computing and Philosophy*. College Park, MD.

Bridewell, W., & Langley, P. (2011). A computational account of everyday abductive inference. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 2289–2294). Austin, TX: Cognitive Science Society.

Briggs, G., & Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. *Proceedings of Twenty-Seventh AAAI Conference on Artificial Intelligence* (pp. 1213–1219). Bellevue, WA: AAAI Press.

de Kleer, J. (1986). An assumption-based truth maintenance system. *Artificial Intelligence*, *28*, 127–162.

Egan, A. (2008). Seeing and believing: perception, belief formation and the divided mind. *Philosophical Studies*, *140*, 47–63.

Gärdenfors, P., & Makinson, D. (1988). Revisions of knowledge systems and epistemic entrenchment. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge* (pp. 83–95). Los Altos, CA: Morgan Kaufmann.

Goldman, A. I. (2006). *Simulating minds: the philosophy, psychology, and neuroscience of mindreading*. New York City, NY: Oxford University Press.

Harmon-Jones, E., & Mills, J. (1999). *Cognitive dissonance: Progress on a pivotal theory in social psychology*. Washington, DC: American Psychological Association.

Hintikka, J. (1962). *Knowledge and belief. An introduction to the logic of the two notions*. Ithaca, NY: Cornell University Press.

Isaac, A. M. C., & Bridewell, W. (in press). Mindreading deception in dialog. *Cognitive Systems Research*.

Konolige, K. (1986). What awareness isn't: A sentential view of implicit and explicit belief. *Proceedings of the First Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 241–250). Monterey, CA: Morgan Kaufmann.

Koons, R. (2013). Defeasible reasoning (supplement on AGM postulates). In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

Kyburg, H. E. (1961). *Probability and the logic of rational belief.* Middletown, CT: Wesleyan University Press.

Lakemeyer, G. (1986). Steps towards a first-order logic of explicit and implicit belief. *Proceedings of the First Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 325–340). Monterey, CA: Morgan Kaufmann.

Levesque, H. J. (1984). A logic of implicit and explicit belief. *Proceedings of the Fourth National Conference on Artificial Intelligence* (pp. 198–202). Austin, TX: AAAI Press.

Makinson, D. (1965). The paradox of the preface. *Analysis*, *25*, 205–207.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An introduction to the syntax and content of Cyc. In *Papers from the 2006 aaai spring symposium on formalizing and compiling background knowledge and its applications to knowledge representation and question answering*, 44–49. Stanford, CA: AAAI Press. Tech. Rep. SS-06-05.

McAllester, D. (1990). Truth maintenance. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 1109–1116). Boston, MA.

Pollock, J. (2008). Defeasible reasoning. In J. E. Adler & L. P. Rips (Eds.), *Reasoning: The study of human inference and its foundations*, 451–470. New York, NY: Cambridge University Press.

Pynadath, D. V., & Marsella, S. C. (2005). PsychSim: Modeling theory of mind with decision-theoretic agents. *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (pp. 1181–1186). Edinburgh, UK: Professional Book Center.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*, 531–553.

Shadyac, T. D., & Grazer, B. P. (1997). Liar liar.

Stalnaker, R. (1993). A note on non-monotonic modal logic. *Artificial Intelligence*, *64*, 183–196.

Stuhlmüller, A., & Goodman, N. D. (in press). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Journal of Cognitive Systems Research*.

Weyhrauch, R. W. (1980). Prolegomena to a theory of mechanized formal reasoning. *Artificial Intelligence*, *13*, 133–170.